

BAYESIAN INFERENCE FOR MATCHED CASE-CONTROL STUDIES

By
SAMIRAN SINHA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2004

Copyright 2004

by

Samiran Sinha

I dedicate this work to my parents.

ACKNOWLEDGMENTS

I would like to take this opportunity to express my sincere gratitude to my two advisors (Professor Malay Ghosh and Professor Bhramar Mukherjee) for their immense help at every stage of my research. I remain grateful for their constant encouragement, and mental support during the hardship of my graduate study at the University of Florida. Despite being a very distinguished, and prolific statistician, Professor Ghosh always found time to listen to my problems. Dr. Mukherjee's advice and support has helped me to face academic as well as nonacademic challenges.

I would also like to thank my other committee members (Professor Alan Agresti, Professor Andrew Rosalsky, and Professor David Wilson) for their time and thoughtful comments on my dissertation. I would like to convey my special thanks to Professor Andrew Rosalsky for many encouraging and fruitful discussions. I truly appreciate the help from Professor Lawrence Winner regarding the writing of SAS programs. I thank all the other Professors in our department for their help throughout my graduate study.

Next, I would like to thank my friends and peers in the department. I thank Ludwig Heigenhauser (who is now in Munich, Germany) for sharing his expertise in R, and LaTeX with me. I thank Dr. Jeff Thompson (currently on the faculty at the North Carolina State University) for very helpful interview tips. I thank my friends Damaris, David, Dobrin, Jie, Keunbaik, Lee, and Pavlina. I thank my friends Ananya, Karabi, Rajarshi, Siuli, Soumita, Sounak, Sourish, Upasana, and Vivekananda for the fun time I spent with them, and for all the rides they

gave me after late-night parties. I thank my friends outside the department whose inspiration boosted my confidence.

Last, but not the least, I would like to thank my family for their unconditional support. I acknowledge the constant emotional support and encouragement I received from my mother, father, and sister. They always dreamt about my future, and helped me in any way they could. Without their presence, I could not have completed my doctoral work. Finally, I thank God for all His help.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	x
 CHAPTER	
1 SELECTED REVIEW OF CASE-CONTROL STUDIES	1
1.1 Introduction	1
1.2 Inferences from 2×2 Table	3
1.3 Test for Homogeneity of Odds Ratio	5
1.4 Matched Pairs	6
1.5 Logistic Regression in Case-Control Studies	7
1.6 Bayesian Analysis of Case-Control Studies	12
1.7 Topics of this Dissertation	20
2 SEMIPARAMETRIC BAYESIAN ANALYSIS OF MATCHED CASE-CONTROL STUDIES WITH MISSING EXPOSURE	22
2.1 Introduction	22
2.2 Model and Notation	25
2.3 Likelihood, Prior and Posterior	27
2.4 Computational Details	32
2.5 Two Special Cases with Data Examples	36
2.5.1 Continuous Exposure: Equine Epidemiology Example	36
2.5.2 Binary Exposure: Endometrial Cancer	40
2.6 Simulation Study	42
2.7 Concluding Remarks	45
3 BAYESIAN SEMIPARAMETRIC MODELLING FOR MATCHED CASE-CONTROL STUDIES WITH MULTIPLE DISEASE STATES	51
3.1 Introduction	51
3.2 Model and Notation	53
3.3 Likelihood, Priors and Posteriors	55
3.4 Example and Computing Scheme	59
3.5 Simulation Study	64

3.6	Concluding Remarks	66
4	SEMIPARAMETRIC BAYESIAN ANALYSIS OF MATCHED CASE-CONTROL STUDIES WHEN EXPOSURE VARIABLES ARE MEASURED WITH ERROR	73
4.1	Introduction	73
4.2	Model and Notation	75
4.3	Likelihood, Priors, and Posteriors	77
4.4	Example: Colon Cancer	80
4.5	Computational Details	82
4.6	Concluding Remarks	85
5	MODELLING ASSOCIATION AMONG MULTIVARIATE EXPOSURES IN CASE-CONTROL STUDIES	88
5.1	Introduction	88
5.2	Model and Notation	91
5.2.1	Bivariate Binary Exposure Variable	94
5.2.2	One Binary Exposure and Another Belonging to an Expo- nential Family	97
5.3	Simulation Study	101
5.4	Concluding Remarks	103
6	SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH	109
	REFERENCES	112
	BIOGRAPHICAL SKETCH	118

LIST OF TABLES

<u>Table</u>	<u>page</u>
1-1 Case-control data with a binary exposure variable	2
1-2 Series of 2×2 table for case-control data	4
1-3 1:1 matched case-control data with a binary exposure variable	6
2-1 Results of the equine data example	47
2-2 Results of the endometrial cancer data example	48
2-3 Results of the simulation study for full data	49
2-4 Results of the simulation study for missing data	50
3-1 Analysis of low-birth-weight data using full dataset with multiple disease states	68
3-2 Analysis of low-birth-weight data after deleting 40% observations on smoking completely at random	69
3-3 Analysis of low-birth-weight data after collapsing categories 1 and 2 into a single low-birth-weight category (less than 2500 grams.)	70
3-4 Results of the simulation study for multiple disease states, part-I . . .	71
3-5 Results of the simulation study for multiple disease states, part-II . . .	72
4-1 Analysis of colon cancer data using full dataset	86
4-2 Analysis of colon cancer data with 10% artificial missing observation in each of the exposure variables	87
5-1 Results of the endometrial cancer data for the association modelling .	105
5-2 Secondary model parameter estimates for endometrial cancer data . .	106
5-3 Results of the benign breast disease data	106
5-4 Results of the simulation study for the associated exposure	107
5-5 Results of the simulation study for the independent exposures	108

LIST OF FIGURES

<u>Figure</u>		<u>page</u>
2-1	Plot of the variability of γ_{0i} 's for the equine data example	38
2-2	Predictive distribution of $\gamma_{0n+\frac{1}{2}}$ for the equine data example	39
2-3	Plot of the variability of γ_{0i} 's for the LA cancer study example	41
2-4	Predictive distribution of $\gamma_{0n+\frac{1}{2}}$ for the LA cancer study example . . .	43
3-1	Plot of the variability of γ_{0i} 's for the low-birth-weight study example.	63

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

BAYESIAN INFERENCE FOR MATCHED CASE-CONTROL STUDIES

By

Samiran Sinha

August 2004

Chair: Malay Ghosh

Cochair: Bhramar Mukherjee

Major Department: Statistics

The classical technique of conditional logistic regression is not efficient for analyzing matched case-control data when some exposure variables have missing observations. I present a semiparametric Bayesian approach to analyze this type of data by modeling the distribution of the exposure variable which may have potentially missing values. It is assumed that the exposure distribution is a member of the exponential family. The novel feature of this method is that it captures unmeasured stratum heterogeneity in the distribution of the exposure variable which is most often ignored or is not explicitly modeled in this context. I use a Dirichlet process prior on the stratum heterogeneity parameters, and parametric prior on all other parameters of our interest, and estimate them through a Markov chain Monte Carlo integration scheme. One attractive feature of this method is its data-adaptive nature for capturing true heterogeneity inherent in the data sets. The method is illustrated by two real life examples.

Next I consider a matched case-control study with multiple disease states. The probability of disease development is described by a multinomial logistic regression model. The goal of the study is to assess the effect of potential risk factor on the

different categories of the disease. As before, I assume that the distribution of the exposure variable belongs to a general exponential family, and allow for possible unmeasured stratum heterogeneity. The proposed method is extended to cases in which measurement error and missingness occur simultaneously in an exposure variable.

Finally I consider association modeling among multiple exposure variables in a matched case-control study when some of the exposure variables may be partially missing. I consider two different types of association models accompanied by two real-data examples. Concluding remarks and some directions for future research are included in the end.

CHAPTER 1 SELECTED REVIEW OF CASE-CONTROL STUDIES

1.1 Introduction

The goal of an epidemiologic study is to find causes of a disease and then assess the degree of association between the disease and its potential risk factors. To collect data for an epidemiologic study, one can use a retrospective study design, or a prospective study design, or a cross-sectional study design, or a combination of all three study designs depending upon the need of the study. The data mainly consist of a disease outcome, and some exposure variables along with some demographic information of subjects. Case-control study designs are retrospective designs, where one first observes the disease outcome of a subject and then the subject is followed back to get exposure information. Thus, a case-control study consists of a group of diseased subjects (cases) and a group of nondiseased subjects (controls) along with their exposure information. Cases and controls can be randomly chosen from a population, or by using stratified random sampling (Scott and Wild, 1986). The present Chapter basically deals with the historical development of how to analyze case-control data with exposure variables, which may be binary, categorical or continuous.

Quite often dataset contains missing observations on some exposure variables. The complete-case analysis which naively excludes subjects with missing exposure values, produces inefficient estimates of the parameters of interest compared to the one which considers all the subjects even if some subjects may have some missing observations. In the following Chapters I propose a new method for analyzing matched case-control data with partially missing observations on exposure variables.

In a matched case-control study, controls are matched with a case (or several cases) on the basis of some matching factors (called confounding variables) such as age, gender, hospital wards etc. Matching is important if exposure variable is strongly associated with matching variables; otherwise unmatched and matched analysis of a matched case-control data will produce comparable estimates of the parameters of our interest.

Consider a case-control study with a binary exposure variable X and disease indicator D . Suppose one observes n_1 cases, and n_0 controls, and they are classified according to a single binary exposure variable X ($X = 1$ exposed, and $X = 0$ unexposed); so from the sample, the exposure odds ratio of cases to controls is

$$\frac{P(X = 1|D = 1)P(X = 0|D = 0)}{P(X = 1|D = 0)P(X = 0|D = 1)} = \frac{n_{11}n_{00}}{n_{10}n_{01}}. \quad (1.1)$$

where n_{11} , and n_{10} be the number of exposed, and unexposed cases while n_{01} and n_{00} be the number of exposed and unexposed controls in the sample. Cornfield

Table 1-1: Case-control data with a binary exposure variable

Disease Status	Exposed	Not Exposed	Total
Case	n_{11}	n_{10}	n_1
Control	n_{01}	n_{00}	n_0
Total	e_1	e_0	N

(1951) showed that disease odds ratio with exposure $X = 1$, relative to that for $X = 0$ is the same as the exposure odds ratio for a diseased person to a nondiseased person. That is,

$$\frac{P(D = 1|X = 1)P(D = 0|X = 0)}{P(D = 0|X = 1)P(D = 1|X = 0)} = \frac{P(X = 1|D = 1)P(X = 0|D = 0)}{P(X = 1|D = 0)P(X = 0|D = 1)}.$$

For a rare disease, both $P(D = 0|X = 0)$, and $P(D = 0|X = 1)$ are approximately equal to 1; so the disease odds ratio, say, $\theta(=\exp(\beta))$, say, where β denotes the

log-odds ratio parameter), becomes the relative risk. That is

$$\theta = \frac{P(D=1|X=1)P(D=0|X=0)}{P(D=0|X=1)P(D=1|X=0)} \approx \frac{P(D=1|X=1)}{P(D=1|X=0)}. \quad (1.2)$$

Note that

$$\begin{aligned} \theta = 1 &\implies \frac{P(D=1|X=1)}{P(D=0|X=1)} = \frac{P(D=1|X=0)}{P(D=0|X=0)} \\ \text{or, } &P(D=0|X=1) = P(D=0|X=0). \end{aligned}$$

So, an odds ratio (relative risk for rare disease) of 1 implies that there is no association (i.e., independence) between the disease and the potential risk factor; whereas an odds ratio other than 1 implies that exposure is either synergistic or antagonistic. It is easy to show that for a large sample, a variance estimate of $\log \hat{\theta}$ is $(1/n_{11} + 1/n_{10} + 1/n_{01} + 1/n_{00})$, provided that none of the cell frequencies is zero. To see this, let $\boldsymbol{\pi} = (\pi_1(1), \pi_0(1))^T$, where $\pi_r(1) = P(X=1|D=r)$ and $\pi_r(1) \in (0, 1)$ for $r = 0, 1$ so $\theta(\boldsymbol{\pi}) = \pi_1(1)\pi_0(0)/\pi_1(0)\pi_0(1)$. Using the Delta method,

$$\begin{aligned} \log(\theta(\hat{\boldsymbol{\pi}})) &= \log(\theta(\boldsymbol{\pi})) + \frac{\partial}{\partial \boldsymbol{\pi}} \log \hat{\theta} |_{\boldsymbol{\pi}=\boldsymbol{\pi}} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) + o_p\left(\frac{1}{\sqrt{N}}\right) \\ &= \log \theta(\boldsymbol{\pi}) + \left\{ \frac{1}{\pi_1(1)(1-\pi_1(1))} + \frac{1}{\pi_0(1)(1-\pi_0(1))} \right\} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \\ &\quad + o_p\left(\frac{1}{\sqrt{N}}\right). \end{aligned}$$

For large samples, $\log(\hat{\theta}) - \log(\theta)$ has an asymptotic normal distribution, with mean 0 and variance $(1/n_{11} + 1/n_{10} + 1/n_{01} + 1/n_{00})$.

1.2 Inferences from 2×2 Table

For a small sample size, inference is based on the conditional distribution of paired binomial data given the marginal totals (which is considered as *ancillary* in the sense that they do not contain any information about the parameter of

interest). So

$$Pr(n_{11}|n_1, n_0, e_1, e_0; \theta) = \frac{\binom{n_1}{n_{11}} \binom{n_0}{e_1 - n_{11}} \theta^{n_{11}}}{\sum_u \binom{n_1}{u} \binom{n_0}{e_1 - n_{11}} \theta^u}. \quad (1.3)$$

The above distribution is called a *non central hypergeometric distribution*. To test $H_0: \theta = \theta_0$ against $K: \theta > \theta_0$, one calculates the upper tail probability under the distribution shown in Equation (1.3),

$$p_u = \sum_{u \geq n_{11}} P(u|n_1, n_0, e_1, e_0; \theta_0) \quad (1.4)$$

which measures the degree of support for H_0 . Similarly to test H_0 against $K: \theta < \theta_0$ we should calculate lower tail probability. The above test is known as *Fisher's exact test*. Altham (1969) gave a Bayesian interpretation of (1.3).

Table 1-2: Series of 2×2 table for case-control data

Disease Status	Exposed	Not Exposed	Total
Case	n_{11i}	n_{10i}	n_{1i}
Control	n_{01i}	n_{00i}	n_{0i}
Total	e_{1i}	e_{0i}	N_i

Suppose now we have a series of 2×2 tables, and the odds ratio is same (say θ) across the series, and we want to test $H_0: \theta = 1$. Under H_0 , the expectation and variance of cell count for exposed cases in the i^{th} table will be $E(n_{11i}|\theta = 1) = n_{1i}e_{1i}/N_i$, $Var(n_{11i}; \theta = 1) = (e_{1i}e_{0i}n_{1i}n_{0i})/N_i^2(N_i - 1)$, and the test statistic (Cochran 1954, Mantel and Haenszel 1959) is

$$\chi^2 = \frac{\left\{ \left| \sum_{i=1}^J (n_{11i} - E(n_{11i}|\theta = 1)) \right| - \frac{1}{2} \right\}^2}{\sum_{i=1}^J Var(n_{11i}|\theta = 1)}. \quad (1.5)$$

1.3 Test for Homogeneity of Odds Ratio

Mantel and Haenszel (1959) proposed a summary odds ratio estimate (relative risk) across the tables

$$\hat{\theta}_{mh} = \frac{\sum_{i=1}^I n_{11i}n_{00i}/N_i}{\sum_{i=1}^I n_{01i}n_{10i}/N_i}. \quad (1.6)$$

The statistic given in (1.5) is unaffected by zero cell counts, and consistently estimates the common odds ratio. In a similar way, one can also test for homogeneity of odds ratios across the tables. One calculates the expected frequency and variance of exposed cases under $\theta = \hat{\theta}_{mh}$, and then construct the test statistic as $\sum_{i=1}^I \{n_{11i} - E(n_{11i}|\hat{\theta}_{mh})\}^2 / \text{Var}(n_{11i}|\hat{\theta}_{mh})$. Under H_0 , the above statistic has an approximate χ^2 distribution with $I - 1$ degrees of freedom. The above formulation could be extended to $K(\geq 2)$ exposure levels (Armitage and Berry, 1994) which was further extended to incorporate stratification. Mantel and Haenszel did not provide any variance estimator for their proposed odds ratio estimator. Robins, Breslow and Greenland (1986) and Phillips and Holland (1987) proposed variance estimator covering two different situations: (i) a large number of tables with small frequencies, and (ii) a small number of tables with large frequencies. The key observation is $E(R_i) = \theta_i E(S_i)$, where $R_i = n_{11i}n_{00i}/N_i$, $S_i = n_{01i}n_{10i}/N_i$, where θ_i is the odds ratio for table 1-2. Assuming a common value of θ_i , say θ , $\hat{\theta}_{mh}$ is the solution of the unbiased estimating equation $R - \theta S = 0$, where $R = \sum_{i=1}^I R_i$, and $S = \sum_{i=1}^I S_i$. Now with one step Taylor's expansion,

$$\begin{aligned} \log(\hat{\theta}_{mh}) &= \log(\theta) + \frac{R - \theta S}{\theta S} + o_p\left(\frac{\text{Var}(R)}{E^2(R)} + \frac{\text{Var}(S)}{E^2(S)}\right) \\ &\doteq \log(\theta) + \frac{R - \theta S}{\theta E(S)} + o_p\left(\frac{\text{Var}(R)}{E^2(R)} + \frac{\text{Var}(S)}{E^2(S)}\right) \\ &= \log(\theta) + \frac{R - \theta S}{E(R)} + o_p\left(\frac{\text{Var}(R)}{E^2(R)} + \frac{\text{Var}(S)}{E^2(S)}\right). \end{aligned}$$

Under paired binomial sampling, the variance of the individual contribution to this estimating equation satisfies

$$N_i^2 \text{Var}(R_i - \theta S_i) = \frac{1}{2} E[\{n_{11i}n_{00i} + \theta n_{01i}n_{10i}\} \\ \times \{n_{11i} + n_{00i} + \theta(n_{01i} + n_{10i})\}].$$

Combining the above two results, one gets

$$\text{Var}(\log(\hat{\theta}_{mh})) \doteq \frac{\text{Var}(R - \theta S)}{E^2(R)} = \frac{\sum_{i=1}^I \text{Var}(R_i - \theta S_i)}{E^2(R)}.$$

Hence, $\hat{\text{Var}}(\log(\hat{\theta}_{mh})) \doteq \sum_{i=1}^I \hat{\text{Var}}(R_i - \theta S_i)/R^2$.

1.4 Matched Pairs

I have already mentioned the role of matching for controlling confounding factors. To avoid gross imbalance in case-control ratio across strata, matching could be introduced at the design stage of a study. The simplest situation of matched data arises when one case is matched with a control, and they are categorized on the basis of a binary exposure.

Table 1-3: 1:1 matched case-control data with a binary exposure variable

Case	Control	
	Exposed	Unexposed
Exposed	m_{11}	m_{10}
Unexposed	m_{01}	m_{00}

Suppose one has m_{11} , m_{10} , m_{01} , and m_{00} matched pairs under different categories. Let π be the probability of observing a matched pair with an exposed case and unexposed control given a discordant pair. Then

$$\pi = \frac{P(X = 1|D = 1)P(X = 0|D = 0)}{P(X = 1|D = 1)P(X = 0|D = 0) + P(X = 0|D = 1)P(X = 1|D = 0)} = \frac{\theta}{\theta + 1}. \quad (1.7)$$

The conditional distribution of m_{10} given the total number of discordant pairs is binomial with parameter $(m_{10} + m_{01}, \pi)$. So the MLE of θ is m_{10}/m_{01} which is

also the Mantel Haenszel estimator of the common odds ratio parameter. Note that when $\theta = 1$, $\pi = 1/2$. Hence the test statistic to test $H_0 : \theta = 1$ is

$$\chi^2 = \frac{(|m_{10} - E_{H_0}(m_{10}|m_{10} + m_{01})| - \frac{1}{2})^2}{\text{var}_{H_0}(m_{10}|m_{10} + m_{01})}. \quad (1.8)$$

This is known as McNemar's (1947) test. One of the potential problem of this estimator, and this test is that it uses only the discordant pairs of observation and discards the concordant set.

In case of 1 : M matching, the Mantel-Haenszel estimator of common odds ratio is

$$\hat{\theta}_{mh} = \frac{\sum_{r=1}^M (M - r + 1)m_{1r-1}}{\sum_{r=1}^M r m_{0r}}, \quad (1.9)$$

where m_{1s} is the number of matched sets where the case and s controls are exposed; and m_{0s} is the number of matched sets where the case is unexposed but s controls are exposed. The test statistic for testing $H_0 : \theta = 1$ is

$$\chi^2 = \frac{\left\{ \left| \sum_{r=1}^M (m_{1r-1} - \frac{r t_r}{M+1}) \right| - \frac{1}{2} \right\}^2}{\sum_{r=1}^M r t_r \frac{(M-r+1)}{(M+1)^2}}, \quad (1.10)$$

where $t_r = m_{1r} + m_{0r}$.

The log-odds ratio regression concerns the effect of a binary exposure on the risk of a disease. Multiple categorical risk factors could be considered, but only by stratifying with respect to the other factors. Methods for considering simultaneous effects of several factors were initiated in 1961 with a linear logistic model (Cornfield et al., 1961). Later Cox (1989) developed the theory of logistic regression in more detail.

1.5 Logistic Regression in Case-Control Studies

Suppose that the covariate $\mathbf{z}$ is an arbitrary p -vector with $\mathbf{z}_0$ as reference category. An odds ratio model

$$P(D = 1|\mathbf{z})P(D = 0|\mathbf{z}_0)\{P(D = 0|\mathbf{z})P(D = 1|\mathbf{z}_0)\}^{-1} = \exp\{\beta^T(\mathbf{z} - \mathbf{z}_0)\} \quad (1.11)$$

is equivalent to a binary logistic failure probability model

$$P(D = 1|\mathbf{z}) = H\{(\alpha + \beta^T(\mathbf{z} - \mathbf{z}_0))\} \quad (1.12)$$

(Prentice and Pike, 1979), where $H(x) = 1/(1 + e^{-x})$ is the logistic distribution function, and $\alpha = \log\{Pr(D = 1|\mathbf{z}_0)/Pr(D = 0|\mathbf{z}_0)\}$ is the log-odds for the disease given the reference value $\mathbf{z}_0$. The covariate density model also takes the logistic form, i.e.,

$$P(\mathbf{z}|D = 1) = P(\mathbf{z}_0|D = 1) \exp[\log\{\frac{P(\mathbf{z}|D = 0)}{P(\mathbf{z}_0|D = 0)}\}^\dagger \beta^T(\mathbf{z} - \mathbf{z}_0)]. \quad (1.13)$$

Let us again consider the unmatched scenario of n_1 cases and n_0 controls, ($n = n_0 + n_1$) but with an arbitrary exposure variable $\mathbf{z}$. The retrospective likelihood function for this unmatched design is

$$L = \prod_{j:\text{cases}} P(\mathbf{z}_j|D = 1) \times \prod_{j:\text{controls}} P(\mathbf{z}_j|D = 0). \quad (1.14)$$

Let S be a sampling indicator, which indicates whether an individual is selected in the case-control sample ($S = 1$) or not ($S = 0$). Since, conditional on disease status, sampling is independent of exposure, using Bayes' theorem,

$$\begin{aligned} P(\mathbf{z}|D = i) &= P(\mathbf{z}|D = i, S = 1) \\ &= \frac{P(D = i|\mathbf{z}, S = 1)P(\mathbf{z}|S = 1)}{P(D = i|S = 1)} \\ &= P(D = i|\mathbf{z}, S = 1)P(\mathbf{z}|S = 1) \frac{n}{n_i}. \end{aligned} \quad (1.15)$$

As in Mantel(1973), one can obtain the conditional probability of a person being a case, given he has exposure $\mathbf{z}$, and that he was sampled for the case-control study as

$$P(D = 1|\mathbf{z}, S = 1) = \frac{P(S = 1|D = 1)P(D = 1|\mathbf{z})}{\sum_{i=0}^1 P(S = 1|D = i)P(D = i|\mathbf{z})}. \quad (1.16)$$

Here I used the fact that sampling is independent of exposure within cases and controls. Inserting (1.12) into (1.16), and using the relationship $Pr(D = 1|S = 1)/Pr(D = 0|S = 1) = n_1/n_0$ one obtains

$$P(D = 1|z, S = 1) = H\{\delta + \beta^T(z - z_0)\}, \quad (1.17)$$

where

$$\delta = \log\left\{\frac{n_1 P(D = 0) P(D = 1|z_0)}{n_0 P(D = 1) P(D = 0|z_0)}\right\}. \quad (1.18)$$

Note that $\delta - \alpha = \log(n_1/n_0) - \log\{P(D = 1)/P(D = 0)\}$, which is the log odds ratio of the disease probabilities in the sample to those in the entire population.

So far I have shown that, starting with a prospective logistic regression model for disease incidence during a defined time interval, one again arrives at a logistic regression model for the retrospective probability of drawing a case from case-control sample. The log-odds ratio parameter β , is the same in both formulations, while the intercept parameters δ , and α are different. While α depends only on disease incidence for the reference value z_0 , δ is also a function of the mean incidence in the population, and of the ratio of cases and controls in the study sample.

Substituting (1.15) in L one obtains

$$L \propto L_1 \times L_2,$$

with the constraint $n_i/n = \int P(z|S = 1)P(D = i|Z, S = 1)dz$, where

$$L_1 = \prod_{j:\text{cases}} P(D = 1|z_j, S = 1) \times \prod_{j:\text{controls}} P(D = 0|z_j, S = 1). \quad (1.19)$$

and

$$L_2 = \prod_{d=0}^1 \prod_{j=1}^{n_d} P(z_{dj}|S = 1)$$

The maximum likelihood estimates (MLE) $\hat{\beta}$ of the log-odds ratio parameter obtained by maximizing L would be the same as the MLE obtained from L_1 (Prentice and Pyke, 1979). It can be shown that the MLE obtained from L_1 is consistent for the log-odds ratio parameter is asymptotically normal.

So far I discussed logistic regression in the context of unmatched studies. Next focus on logistic regression in the matched case-control setup. In the simplest setting, the data consist of s strata and there are M_i controls matched with a case, for stratum $S_i, i = 1, \dots, s$. As before, I assume a prospective logistic incidence model for disease

$$P(D = 1 | \mathbf{z}, S_i) = H\{\alpha_i + \beta^T(\mathbf{z} - \mathbf{z}_0)\}, \quad (1.20)$$

where α_i 's are stratum specific intercept term. Proceeding as in the unstratified case, the retrospective logistic model becomes

$$P(D = 1 | \mathbf{z}, S_i, S = 1) = H\{\beta_0(S_i) + \beta^T(\mathbf{z} - \mathbf{z}_0)\}, \quad (1.21)$$

where

$$\beta_0(S_i) = \log\left\{\frac{1 \times P(D = 0 | S_i)}{M_i \times P(D = 1 | S_i)} \cdot \frac{P(D = 1 | \mathbf{z}_0, S_i)}{P(D = 0 | \mathbf{z}_0, S_i)}\right\}. \quad (1.22)$$

Assuming that the 1st subject in each stratum is a case and rest of the subjects are controls, the likelihood under this retrospective sampling is

$$\begin{aligned} L &= \prod_{i=1}^s \{P(\mathbf{z}_{i1} | D = 1, S_i) \prod_{j=1}^{M_i} P(\mathbf{z}_{ij} | D = 0, S_i)\} \\ &\propto \prod_{i=1}^s \{P(D = 1 | \mathbf{z}_{i1}, S_i, S = 1) \prod_{j=1}^{M_i} P(D = 0 | \mathbf{z}_{ij}, S_i, S = 1)\}. \end{aligned} \quad (1.23)$$

The above likelihood involves β and a set of nuisance parameters $\beta_0(S_i)$ for $i = 1, \dots, s$ which grows in direct proportion to the number of strata. This gives rise to the well known Neyman-Scott phenomenon where the MLE turns out to be inconsistent. Conditioning on the sufficient statistics $\sum_{j=1}^{M_i+1} D_{ij}$ of α_i , one obtains

the conditional likelihood

$$\begin{aligned}
 L_c &= \prod_{i=1}^s \prod_{j=1}^{M_i+1} P(D_{ij} | \mathbf{z}_{ij}, S_i, S = 1, \sum_{j=1}^{M_i+1} D_{ij} = 1) \\
 &= \prod_{i=1}^s \frac{\exp(\beta^T \mathbf{z}_{i1})}{\sum_{j=1}^{M_i+1} \exp(\beta^T \mathbf{z}_{ij})}, \tag{1.24}
 \end{aligned}$$

assuming that in each stratum $j = 1$ stands for case. The above method is known as conditional logistic regression (CLR).

The advantage of the above likelihood is (a) it involves only the relative risk parameters, and (b) one obtains the same form of likelihood as either (i) the data arising from a prospective study of $\sum_{i=1}^s M_i + s$ individuals, the conditioning event being the observed number s of cases arising in the sample; or (ii) case-control study involving s cases and $\sum_{i=1}^s M_i$ controls, the conditioning event being $\sum_{i=1}^s M_i + s$ exposure histories.

Note that if $\mathbf{x}$ is some matching variable, then its contribution to the conditional likelihood is zero; so one can not estimate the relative risk corresponding to the matching variable. One can consider an interaction term of a matching variable with other covariates, and investigate the change of the relative risk parameter across strata (Greenland and Robins, 1985).

Prentice and Breslow (1978) gave a conceptual foundation for case-control study deriving conditional likelihood(CL) from failure time consideration. One of the main drawback of CL is that, if there is some missing information on an exposure variable of a subject, the available partial information on that subject is ignored from the CLR. For a 1 : 1 matched case-control study, missing observation on an exposure of a subject leads to deletion of the entire stratum from the CLR. In that situation, one only uses the strata which have complete information for all the subjects, and lose information on strata containing any missingness. Substantial amount of work is available on the issue of missing exposure. There

are two basic approaches to handle this problem. One is to model the distribution of exposure variable and the other is to model the missingness process. Lawless et al. (1999) proposed a semiparametric method to deal with the exposure distribution. Paik and Sacco (2000) modeled the distribution of exposure variables parametrically, and then used pseudo conditional likelihood method to estimate the parameters. Lin and Ying (1993) considered missing covariate values in Cox's model. Other studies address the issue of misclassification of exposure variables and mismeasured covariates (Carroll et al., 1993; Carroll et al., 1995; Dahm et al., 1995), as well as sample size determination.

1.6 Bayesian Analysis of Case-Control Studies

So far we discussed case-control studies in frequentist framework. Now I focus on Bayesian analysis of case-control problems. Zelen and Parker (1986), Nurminen & Mutanen (1987), Marshall (1988), and Ashby et al. (1993) addressed the Bayesian method for case-control study with only a single binary exposure variable. All of them used the following model for a binary exposure variable.

Let ϕ and γ be the probabilities of exposure in control and case populations respectively. The retrospective likelihood is

$$l(\phi, \gamma) \propto \phi^{n_{01}}(1 - \phi)^{n_{00}}\gamma^{n_{11}}(1 - \gamma)^{n_{10}}, \quad (1.25)$$

where n_{01} and n_{00} are the number of exposed and unexposed observations in a control population; whereas n_{11} and n_{10} denote the same for a case population.

Independent conjugate prior distributions for ϕ and γ are assumed to be $Beta(u_1, u_2)$ and $Beta(v_1, v_2)$ respectively. After reparametrization, one obtains the posterior distribution of the log odds ratio parameter, $\beta = \log\{\gamma(1 - \phi)/\phi(1 - \gamma)\}$

as

$$p(\beta|n_{11}, n_{10}, n_{01}, n_{00}) \propto \exp\{(n_{11} + v_1)\beta\} \\ \times \int_0^1 \frac{\phi^{n_{11}+n_{01}+v_1+u_2-1}(1-\phi)^{n_{10}+n_{00}+v_2+u_1-1}}{\{1-\phi+\phi\exp(\beta)\}^{n_{11}+n_{10}+v_1+v_2}} d\phi \quad (1.26)$$

The posterior density of β does not exist in closed form, but may be evaluated by numerical integration.

Since interest often lies in the hypothesis $\beta = 0$, Zelen and Parker (1986) recommend calculating the ratio of the two posterior probabilities $p(\beta)/p(0)$ at selected deviates β . When β is set at the posterior mode, a large value of this ratio will indicate concentration of the posterior away from 0 and one would infer disease-exposure association. However the critical value suggested for this ratio is completely arbitrary. They also provide a normal approximation to the posterior distribution of β to avoid numerical computation. Zelen and Parker (1986) also discuss the problem of choosing a prior distribution based on some prior data on exposure information. They also raise the possibility of performing a case-control analysis without the presence of a control sample, if adequate exposure information is available from prior studies.

Nurminen and Mutanen (1987) considered a more general parameterization in terms of the odds ratio $R = \exp(\beta)$ which covers risk ratio and risk differences. They provided a complicated exact formula for the cumulative density function of this general comparative parameter. The formula can be related to Fisher's exact test for comparing two proportions in sampling theory. The Bayesian point estimates are considered as posterior median, and mode, whereas inference is based on highest posterior density interval for the comparative parameter of interest.

Marshall (1988) provided a closed-form expression for the moments of the posterior distribution of the odds ratio. He mentioned that an approximation to the exact posterior density of the odds ratio parameter can be obtained by power

series expansion of the hypergeometric functions involved in the expression for the density but acknowledges the problem of slow convergence in adopting this method. Instead Marshall used Lindley's (1964) result for the approximate normality of $\log(R)$ which works very well over a wide range of situations. In the absence of exposure information, Marshall recommended using independent priors on the parameters. He suggested that a perception about the value of the odds ratio should guide the choice of prior parameters rather than attempting to exploit the exposure proportions as suggested in Zelen-Parker. Inference again is based on posterior credible intervals.

Müller and Roeder (1997) presented a semiparametric Bayesian approach to case-control studies having continuous exposures with measurement error. The outline of their approach is as follows:

Let D be a binary response observed together with covariates X , and Z , where Z may be vector-valued. The retrospective case-control sample is chosen by selecting n_1 cases with $D = 1$ and n_0 controls with $D = 0$. The error in variables occurs when for a subset of the data—reduced data or ($\mathcal{R}$) from now on—a surrogate W is measured instead of the true covariate X . For a typical validation study, for the remaining subjects—complete data ($\mathcal{C}$) from now on—one has observations on both the gold standard X and the surrogate W .

The approach proposed by Müller and Roeder (1997) is based on a nonparametric model for the retrospective likelihood of the covariates, and exposure. For reduced data, this nonparametric distribution models the joint distribution of (Z, W) and the missing covariate X_R . For complete data, the same distribution models the joint distribution of the observed covariates (X_C, Z, W) . The nonparametric distribution is a class of flexible mixture distributions, obtained by using mixture of normal models with a Dirichlet Process prior on the mixing measure (Escobar and West, 1995). For the prospective disease model, relating disease to

exposure is taken to be a logistic distribution characterized by a vector of parameters β . Also, let θ denote the parameters describing the marginal distribution of the exposure and covariates. Under certain conditional independence type assumptions and nondifferential measurement error, the retrospective likelihood is shown to be compatible with the prospective model. With a prior $p(\beta, \theta)$, one obtains the joint posterior as,

$$p(\beta, \theta | X_C, W, D, Z) \propto p(\beta, \theta) \prod_{i \in C} p(X_i, Z_i, W_i | D_i, \beta, \theta) \prod_{i \in R} \left\{ \int p(X_i, Z_i, W_i | D_i, \beta, \theta) dX_i \right\} \quad (1.27)$$

By augmenting the parameter with the latent vector, one can write the joint posterior of parameters and missing exposure X_R as

$$p(\beta, \theta, X_R | X_C, W, D, Z) \propto p(\beta, \theta) \prod_{i=1}^n p(X_i, Z_i, W_i | D_i, \beta, \theta). \quad (1.28)$$

The above likelihood can be factorized by using Bayes theorem under the additional assumption of nondifferential measurement error: $p(D | X, Z, W, \beta) = p(D | X, Z, \beta)$, as

$$p(\beta, \theta, X_R | X_C, W, D, Z) \propto p(\beta, \theta) \prod_{i=1}^n p(X_i, Z_i, W_i | \theta) p(D_i | X_i, Z_i, \beta) / p(D_i | \beta, \theta), \quad (1.29)$$

where $p(D_i | \beta, \theta) = \int p(D_i, X_i, Z_i, W_i, \beta) dP(X_i, Z_i, W_i | \theta)$. Let $\theta = (\theta_1, \theta_2, \dots, \theta_n)$.

Then the following mixture model with a multivariate normal kernel ϕ_{θ_i} is assumed:

$$p(X_i, Z_i, W_i | \theta_i) = \phi_{\theta_i}(X_i, Z_i, W_i) \quad (1.30)$$

$$\theta_i \sim G$$

$$G \sim DP(\alpha G_0).$$

Where $\theta_i = (\mu_i, \Sigma_i)$, the parameters of the trivariate normal kernel, and $DP(\alpha G_0)$ denotes a Dirichlet process with a base measure G_0 , and concentration parameter α . The hierarchy is completed by assuming a normal/Wishart hyper-prior on the parameters of the base measure $G_0 = N(\mu_0, \Sigma_0)$, and a Gamma prior on the concentration parameter α . A noninformative prior is assumed on β . A Markov chain Monte Carlo numerical integration scheme is designed for estimating the parameters. The computation scheme becomes either very expensive or inaccurate when the dimension of the (X, Z, W) -space increases as one must then evaluate $p(D_i | \beta, \theta)$ by numerical integration over the (X, Z, W) -space. This paper breaks many new grounds in Bayesian treatment of case-control studies including consideration of continuous covariates; measurement error; and flexible nonparametric modeling for exposure distribution leading to robust, interpretable results quantifying the effect of exposure.

Müller and Roeder point out that categorical covariates require a different treatment, assuming a multivariate normal kernel for the Dirichlet Process implicitly assumes continuous exposures. Seaman and Richardson (2001) extend the binary exposure model of Zelen-Parker to any number of categorical exposures, by simply replacing the binomial likelihoods in (1.25) by a multinomial likelihood, and then adopting a MCMC strategy to estimate the log odds-ratio for the disease in each category with respect to a baseline category. The set of multinomial probabilities corresponding to exposure categories in case and control populations is assumed to have a discrete Dirichlet distribution prior. The prior on the log-odds ratios could be chosen to be anything.

Seaman and Richardson also adapted the Müller-Roeder approach to categorical covariates by simply modeling the multinomial probabilities corresponding to exposure categories in the entire source population as opposed to looking at case and control populations separately as in the generalized Zelen-Parker binary

model described above. A noninformative Dirichlet $(0,0,\dots,0)$ prior is assumed on the exposure probabilities. Seaman and Richardson established the equivalence of the generalized binary and adapted the Müller-Roeder approach in some limiting cases with uninformative Dirichlet $(0,0,\dots,0)$ prior on ϕ .

Continuous covariates were treated in Seaman and Richardson (2001) framework by discretizing them into groups. Little information is lost if discretization is sufficiently fine. This opens up the possibility of treating continuous and categorical covariates simultaneously which is not obvious in the Müller-Roeder approach.

Müller et al. (1999) proposed a hierarchical Bayesian approach for combining case-control study and a prospective cohort study, and to estimate the absolute risk of the disease. They modeled retrospective distribution of the exposure variable given the disease status, and accounted for parameter heterogeneity across studies by using a hierarchical Bayesian approach. A mixture model is assumed for the exposure variable. Parameters were estimated via Markov chain Monte Carlo integration scheme.

Diggle, Morris and Wakefield (2000) presented the first Bayesian analysis for cases individually matched to the controls (resulting nuisance parameters are introduced to represent the separate effect of matching in each matched set). They considered matched data when exposure of primary interest is defined by the spatial location of an individual relative to a point or line source of pollution.

The basic model of Diggle et al. (2000) for an unmatched set-up is as follows. Let locations of individuals in the population at risk follow an inhomogeneous Poisson process with intensity function $g(x)$, and the probability that an individual at location x becomes a case is $p^*(x)$. Then the odds of disease among sampled individuals are

$$r(x) = (a/b)p^*(x)/(1 - p^*(x)),$$

where a and b are the sampled proportions of cases and controls. Further, $r(x)$ is modeled as a nonlinear logistic regression model, namely,

$$r(x) = \rho h(x, \theta)$$

where $h(x, \theta)$ involves interpretable parameters of interest θ . The extension of this model to J individually matched case-control pairs in a study region is made in the following way. Suppose, the J locations for cases and controls are denoted by x_{0j} , and x_{1j} respectively, $j=1, \dots, J$. The probability of disease for an individual at location x in stratum j is modeled as:

$$p_j(x) = \frac{r_j(x)}{1+r_j(x)} = \frac{\rho_j h(x, \theta)}{1+\rho_j h(x, \theta)},$$

where ρ_j are varying baseline odds between matched pairs and are considered as nuisance parameters. Conditioning on the unordered set of exposures within each matched set, one has the conditional probability of disease of an individual at distance x in stratum j

$$p_c(x_{j0}) = \frac{h(x_{j0}, \theta)}{h(x_{j0}, \theta) + h(x_{j1}, \theta)}.$$

In case of 1 : M matching in each stratum, and in the presence of q additional covariates $z_k(x_{ji})$ measured for the i -th individual in j -th stratum at location x_{ji} , $i = 1, \dots, M; j = 1, \dots, J; k = 1, \dots, q$, the full fledged conditional probability can be modeled in the form

$$p_c(x_{j0}) = \frac{h(x_{j0}, \theta) \exp(\sum_{k=1}^q z_k(x_{j0}) \phi_k)}{\sum_{i=0}^M h(x_{ji}, \theta) \exp(\sum_{k=1}^q z_k(x_{ji}) \phi_k)},$$

with the conditional log likelihood being of the form

$$L(\theta, \phi) = \sum_{j=1}^J \log p_c(x_{j0}).$$

Diggle et al. (2000) considered this likelihood as well as Bayesian approach for estimation of parameters and inference in this model with specified form of $f(x, \theta)$. Often in studying environmental pollution, there is a putative point source at location x^* , and interest is on how the risk surface changes in relation to x^* . Let

$d = \|x - x^*\|$ be the distance of location x from the source. In such instances they propose h of the form

$$h(x) = 1 + \alpha \exp(-(d/\beta)^2), \text{ for fixed } x^*.$$

This model has a natural interpretation, with α being the proportional increase in disease odds at the source, while β measures the rate of decay with increasing distance from the source in units of distance. The standard likelihood inference for such models based on the likelihood ratio statistic may not be reliable, as the likelihood involved is highly nonregular. The Bayesian paradigm provides an alternative when likelihood surface is highly irregular. In their paper independent normal priors are assumed on the covariate related regression parameters ϕ , and independent uniform priors are chosen for α and β . Estimation is carried out by adopting an MCMC technique, and inference is conducted by computing the Bayes factor for testing the null hypothesis $h(x) = 1$ which amounts to saying that the risk of the disease does not depend on the distance from the putative source.

Ghosh and Chen (2002) developed general Bayesian inferential techniques for matched case-control problems in the presence of one or more binary exposure variables. Their model was more general than that of Zelen and Parker (1986). Also their analysis, unlike that of Diggle, Morris, and Wakefield (2000), was based on an unconditional likelihood rather than a conditional likelihood after elimination of nuisance parameters. Any methodology based on the conditional likelihood is applicable only for the logit link. The general Bayesian methodology based on the full likelihood as proposed by Ghosh and Chen worked beyond the logit link. Their procedure included not only the probit and the complementary log links but also some new symmetric as well as skewed links. The propriety of posteriors was proved under a very general class of priors that need not always be proper. Also, some robust priors (such as multivariate t-priors) were used. The Bayesian procedure was implemented via Markov chain Monte Carlo.

To illustrate their general methodology, Ghosh and Chen considered the Breslow and Day's (1980) the Los Angeles (LA) cancer data, and studied the effect of gall-bladder disease on the risk of endometrial cancer taken as a single exposure variable; and both gall-bladder disease and hypertension as multiple exposure variables. For situations with one case, and one control, it was shown that the Bayesian procedure could yield conclusions different from those of conditional likelihood regarding association between the exposures and the disease. This is because, while the conditional likelihood ignores all concordant pairs, the Bayesian procedure, based on the full likelihood also uses these pairs.

Seaman and Richardson (2004) showed the equivalence of prospective and retrospective analysis of case-control studies in a Bayesian analysis. They showed that the posterior distribution of the log-odds ratio obtained from a prospective and a retrospective likelihood are equivalent upto a proportionality constant.

1.7 Topics of this Dissertation

Bayesian analysis of case-control study got less attention than frequentist approach. Specifically, there are only a few Bayesian works related to problems of matched case-control study. This dissertation is designed to deal with missing data problem in a matched case-control study. Each matched set consists of a case and several controls; and for each subject in the study, one observes a disease variable D , an exposure variable(s) X , and a covariate Z . X is assumed to be missing at random (MAR). Here I develop a method to handle MAR data which exploits all the available information on a subject even if the information on some of the exposure variables are missing for the subject. Logistic distribution is assumed for the prospective model of the disease status which may vary across strata. A full-likelihood-based approach is used by assuming a parametric distribution of the missing exposure variable among the control population and finally the odds ratio parameters are estimated in a fully Bayesian framework. The major departure of

this research is to allow stratum heterogeneity in the exposure distribution and how one handles this heterogeneity parameters in a fully Bayesian framework by using a Dirichlet process prior with a normal base measure. In case of fully observed data, this assumption of heterogeneity may not be crucial, but in the presence of missing data and unmeasured stratum effect, it makes a lot of difference in the inference. I extend the above method to some orthodox data situations: (i) when the case has multiple categories, (ii) when the exposure variables have measurement error and contain partial missing observations, and (iii) when case-control data contains multiple exposure variables which are associated among themselves. The outline of the rest of the dissertation is as follows. In Chapter 2, I deal with the problem of a missing exposure variable in a matched case-control study. A semiparametric Bayesian method is proposed to handle the missingness, and to capture unexplained stratum heterogeneity on the distribution of the exposure variable. In Chapter 3, I extend the proposed Bayesian semiparametric method to a matched case-control study with missing exposure variable when the disease has multiple categories. A unified semiparametric Bayesian approach is presented in Chapter 4, to handle simultaneous occurrence of measurement error and missingness in an exposure variables. In Chapter 5, I consider the problem of modeling association among multiple exposure variables with partially missing values. Chapter 6 offers concluding remarks and directions for future research.

CHAPTER 2

SEMIPARAMETRIC BAYESIAN ANALYSIS OF MATCHED CASE-CONTROL STUDIES WITH MISSING EXPOSURE

2.1 Introduction

In this Chapter, we consider matched case-control studies when some exposure variables are not observed for all study subjects. For example, in the Los Angeles Study on endometrial cancer in a retirement community (Breslow and Day, 1980), a binary variable denoting obesity was missing in approximately 16% of respondents. So far, there are two main approaches to handle the missingness in matched case-control problems. Lipsitz, Parzen and Ewell (1998), and Rathouz, Satten and Carroll (2002) modeled the missingness process directly. More germane to my approach is the likelihood method that involves modeling the missing covariate distribution among the controls, see Satten and Kupper (1993a, 1993b), who initiated this approach. Wang, Wang and Carroll (1997), and Paik and Sacco (2000) attacked the missing data problem directly by positing that the exposure distribution among the controls belongs to a canonical exponential family: while the latter's model is fully parametric, their estimation method is a pseudolikelihood methodology rather than full likelihood, and in the absence of missing data reduces to the usual conditional logistic regression (CLR). Satten and Carroll (2000) generalized the Satten–Kupper and Paik–Sacco methodology to allow a full parametric likelihood analysis, allowing essentially any exposure distribution among the controls, and providing for full likelihood analysis: their approach does not reduce to the usual CLR with no missing data.

I develop a Bayesian approach to case-control studies with a disease indicator D , a completely observed covariate vector $\mathbf{Z}$, and an exposure or risk factor X

with possibly missing values. I will also include variables $S = (S_o, S_u)$ that define the strata used for matching. These are generally a vector of covariates, so that S_o indicates covariates that are explicit factors making up the strata and S_u are factors that are not explicitly used to form strata but nonetheless are associated with the strata. The variable X that is partially missing can be discrete or continuous.

An example will help to clarify the strata issue. There has been a recent resurgence of interest in genetic case-control studies to explore a variety of gene-disease, and gene-environment interactions. The prime objective here is to examine the association between a candidate gene and the occurrence of a disease. In such problems, it is very common to stratify a population which comprises subpopulations with different allele frequencies for the gene under consideration. If different subpopulations have different risks for the disease, then the failure to take these stratum effects into account will lead to misleading estimates of the association between the disease and the candidate gene.

A second example, specifically considered in this chapter, is due to Kim, Cohen and Carroll (2002). They described a 1:1 matched case-control study that investigated the association of various management practices with equine colic. Participating veterinarians were asked to provide data monthly for one horse treated for colic and one horse that received emergency treatment for any condition other than colic, between March 1, 1997, and February 28, 1998. A case of colic was defined as the first horse treated during a given month for signs of intra-abdominal pain. A control horse was defined as the next horse that received emergency treatment for any condition other than colic. Two individual-level variables of interest were the age of the horse X and an indicator Z of whether the horse had undergone a recent change in diet.

In this example, it is difficult to model parametrically the “next sick horse in the clinic” matching method by explicit variables S_o . One can certainly try, e.g., by accounting for month, veterinarian, region of Texas, urban/rural, etc., but the dimensionality is likely to be high, and fairly clearly there is more to the matching than can be measured explicitly. However, it is entirely possible that some or all of these factors affect the distribution of X among the controls, and a likelihood analysis would have to account for them. The dimensionality of the effects quickly gets out of hand in my example and it is likely that in such cases the possible effect of stratification on the distribution of X will be ignored.

The likelihood approaches referenced above start with an explicit parametric distribution for X among the controls, $D = 0$, and given $\mathbf{Z}$ as well as possibly the explicit parts S_o of the strata. Wang et al. (1997) and Paik and Sacco (2000) do not include the strata in the model for X , while Satten and Carroll (2000) include the observed components S_0 as a linear term. In the equine example, I of course believe that it is effectively impossible to measure all the stratification variables, and then to model their effect on X , which is likely to be in the form of an unexplained heterogeneity.

My approach is to generalize the previous methods to allow for the possibility of unmeasured stratification effects, i.e., unexplained heterogeneity in the distribution of X given $\mathbf{Z}$ among the controls. Alternatively, my approach can be looked at as a way of allowing for high dimensional effects of stratification. The stratification effects is modeled via a Dirichlet process prior with a normal base measure for the distribution of the stratum effects, and then employ the Bayesian machinery. By this route, one achieves a measure of model robustness, and acknowledge that the distribution of X among the controls can be affected by stratification, and especially by unmeasured factors.

The outline of this Chapter is as follows. Section 2.2 introduces the model, and notation, while Section 2.3 derives the appropriate likelihoods and the MCMC method for computing Bayesian inference. One notes in passing that a limiting case of my approach is a full parametric Bayesian analysis of matched case-control studies with missing data. Section 2.4 contains computational details of the proposed method. In Section 2.5, I provide data analysis for two examples covering two special cases of the general exponential family: (i) a continuous exposure in the equine epidemiology problem and (ii) a binary exposure in an endometrial cancer study. Section 2.6 contains a small simulation study to assess the performance of our proposed methods. I show that at least for this simulation, the proposed methods lose nothing in the way of efficiency when there are no unmeasured stratification effects on the missing covariate X , but gain quite a bit of efficiency when stratification does affect the missing covariate. Section 2.7 contains concluding remarks.

In concluding this section, I highlight some of the newer aspects of this work. To my knowledge, this is the first attempt which provides a Bayesian approach, both parametric, and semiparametric, towards the analysis of matched case-control studies with missing exposure. Our method is applicable to both discrete, and continuous data. Second, the proposed method captures explicitly the unmeasured stratum effects which may not be possible just by introducing additional covariates.

2.2 Model and Notation

For subject j in stratum i , $i = 1, \dots, n$, $j = 1, \dots, M + 1$, let D_{ij} represent the disease status, namely, $D_{ij} = 1$ for a case $D_{ij} = 0$ for a control. I consider a single exposure variable X_{ij} which may be partially missing and a completely observed p -dimensional vector of covariates $\mathbf{Z}_{ij}$. I assume that the prospective conditional logistic distribution for the disease status is

$$\text{pr}(D_{ij} = 1 | X_{ij}, \mathbf{Z}_{ij}, S_i) = H\{\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2 X_{ij}\}, \quad (2.1)$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here $\beta_0(S_i)$ is a term representing the stratum effect on the probability of belonging to a particular disease state, and β_1 and β_2 are coefficients corresponding to the covariates and exposure in the above logistic regression model. The usual method is to work with the conditional likelihood of the D_{ij} with $\sum_{j=1}^{M+1} D_{ij}$, $i = 1, \dots, n$ as the conditioning variables. In this way, the nuisance parameters $\beta_0(S_i)$ are eliminated.

In the presence of a missing X_{ij} for even one subject, a naive application of the above method leads to removal of the entire i -th stratum even though observations on $(D, \mathbf{Z}, X)$ for other subjects may still be available. This naturally entails a loss of information. To overcome this problem I adopt the following approach.

Let Δ_{ij} represent an indicator variable which assumes the value 0 if the exposure value is missing for subject $j = 1, \dots, M + 1$ in stratum $i = 1, \dots, n$ and is 1 otherwise. The basic structure of the model is

$$p(D_{ij}, X_{ij}, \Delta_{ij} | \mathbf{Z}_{ij}, S_i) = p(X_{ij} | D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}, S_i) \times p(\Delta_{ij} | D_{ij}, \mathbf{Z}_{ij}, S_i) \times p(D_{ij} | \mathbf{Z}_{ij}, S_i). \quad (2.2)$$

Following Satten and Carroll (2000), I will assume that the X 's are missing at random in the sense of Little and Rubin (1987), i.e., $p(X_{ij} | D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}, S_i) = p(X_{ij} | D_{ij}, \mathbf{Z}_{ij}, S_i)$ and that $p(\Delta_{ij} | D_{ij}, \mathbf{Z}_{ij}, S_i)$ does not depend on any parameters of interest. Next I model the distributions of the exposure variable X_{ij} among the controls as

$$p(X_{ij} | D_{ij} = 0, \mathbf{Z}_{ij}, S_i) = \exp [\xi_{ij} \{\theta_{ij} X_{ij} - b(\theta_{ij})\} + c(X_{ij}, \xi_{ij})], \quad (2.3)$$

where θ_{ij} is the natural parameter, and will be modeled as

$$\theta_{ij} = \gamma_{0i} + \boldsymbol{\gamma}^T \mathbf{Z}_{ij}, \quad (2.4)$$

where γ_{0i} , the varying intercept, captures the stratum effect on the natural parameter θ_{ij} , and consequently the stratum effect on the exposure distribution.

Paik and Sacco (2000) adopt a simpler parametric model with $\gamma_{0i} = \gamma_0$. If one writes the stratification variables as $S_i = (S_{0i}, S_{ui})$ with S_{0i} observed and S_{ui} unobserved, Satten and Carroll (2000) set γ_{0i} in (2.4) to be a parametric linear function of S_{0i} . As argued in the introduction, it may not be possible to model the γ_{0i} explicitly in this way as a function of the stratification effects, and my methods differ from others in the fact that I address this issue.

In the following section, I derive the joint distribution for (D_{ij}, X_{ij}) conditional on the sums $\sum_{j=1}^{M+1} D_{ij}$, and using (2.1)-(2.3).

2.3 Likelihood, Prior and Posterior

In this section, I derive the likelihood function, state our prior distribution, and derive the posterior. The key aspect of the modeling is in how one handles the strata effects in the model for the missing covariate among the controls.

In what follows, I will do calculations concerning X as if it were a continuous variable, and hence use integration: if X is discrete the integrals are replaced by sums. The key result is establishing the odds when X is not considered.

Let us denote $\rho(X, \mathbf{Z}, S)$ as the prospective odds of the disease for the exposure X , the completely observed covariate $\mathbf{Z}$, and the stratum S . Now following Satten and Carroll (2000), I have the following results.

Lemma 1.

$$\frac{\text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i)} = \int \rho(X_{ij}, \mathbf{Z}_{ij}, S_i) p(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) d\mu(X_{ij}) \quad (2.5)$$

Lemma 2.

$$p(X_{ij} | D_{ij} = 1, \mathbf{Z}_{ij}, S_i) = \frac{p(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) \rho(X_{ij}, \mathbf{Z}_{ij}, S_i)}{\int \rho(X_{ij}, \mathbf{Z}_{ij}, S_i) p(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) d\mu(X_{ij})}. \quad (2.6)$$

Here $\mu(\cdot)$ is a σ -finite dominating measure. The above two Lemmas hold in full generality and do not need logistic assumption for the prospective disease model.

Notice that from the prospective odds model and the exposure distribution in

the control population, one is able to derive the exposure distribution in the case population. Following are the special cases of the above two Lemmas when the distribution of the exposure variable in the control population belongs to generalized exponential family, as given in (2.3).

Lemma 3.

$$\frac{\text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i)} = \exp [\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \xi_{ij} \{b(\theta_{ij}^*) - b(\theta_{ij})\}], \quad (2.7)$$

where $\theta_{ij}^* = \theta_{ij} + \xi_{ij}^{-1} \beta_2$.

Proof:

$$\begin{aligned} \text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i) &= \int \exp\{\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2 X_{ij}\} \text{pr}(D_{ij} = 0 | X_{ij}, \mathbf{Z}_{ij}, S_i) \\ &\quad \times p(X_{ij} | \mathbf{Z}_{ij}, S_i) dX_{ij} \\ &= \exp\{\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij}\} \int \exp(\beta_2 X_{ij}) p(X_{ij}, D_{ij} = 0 | \mathbf{Z}_{ij}, S_i) dX_{ij} \\ &= \exp\{\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij}\} \text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i) \\ &\quad \times \int \exp(\beta_2 X_{ij}) p(X_{ij} | D_{ij} = 0, \mathbf{Z}_{ij}, S_i) dX_{ij}. \end{aligned}$$

Result (2.7) is important because it means that prospectively, the marginal (over X) probability of an event is logistic, and hence that standard conditional logistic regression calculations can be employed.

Lemma 4.

$$p(X_{ij} | D_{ij} = 1, \mathbf{Z}_{ij}, S_i) = \exp[\xi_{ij} \{\theta_{ij}^* X_{ij} - b(\theta_{ij}^*)\} + c(X_{ij}, \xi_{ij})]. \quad (2.8)$$

Proof:

$$\begin{aligned} p(X_{ij} | D_{ij} = 1, \mathbf{Z}_{ij}, S_i) &= \frac{p(X_{ij}, D_{ij} = 1 | \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i)} \\ &= \frac{\text{pr}(D_{ij} = 1 | X_{ij}, \mathbf{Z}_{ij}, S_i) \times p(X_{ij} | \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i)} \\ &= \frac{\exp(\beta_2 X_{ij}) \text{pr}(D_{ij} = 0, X_{ij} | \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i) \exp[\xi_{ij} \{b(\theta_{ij} + \xi_{ij}^{-1} \beta_2) - b(\theta_{ij})\}]} \end{aligned}$$

$$\begin{aligned}
&= \frac{\exp(\beta_2 X_{ij}) \text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i) p(X_{ij} | D_{ij} = 0, \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i) \exp[\xi_{ij} \{b(\theta_{ij}^*) - b(\theta_{ij}^*)\}]} \\
&= \exp[\xi_{ij} \{\theta_{ij}^* X_{ij} - b(\theta_{ij}^*)\} + c(X_{ij}, \xi_{ij})].
\end{aligned}$$

I assume, without loss of generality, $D_{i1} = 1, D_{i2} = 0, \dots, D_{iM+1} = 0$, for each stratum i . Note that I only need to define a probability model for $[D_{ij} | X_{ij}, \mathbf{Z}_{ij}, S_i]$, and $[X_{ij} | D_{ij} = 0, \mathbf{Z}_{ij}, S_i]$ to arrive at the joint conditional likelihood of $[D_{ij}, X_{ij} | \mathbf{Z}_{ij}, S_i, \sum_{j=1}^M D_{ij} = 1, \Delta_{ij}]$. The expression for this likelihood is based on the following key factorization:

$$\begin{aligned}
L_c &\propto \prod_{i=1}^n \text{pr}\{(D_{i1} = 1, D_{ij} = 0, j \neq 1), X_{ij} | S_i, \mathbf{Z}_{ij}, \Delta_{ij}, \sum_{j=1}^{M+1} D_{ij} = 1\} \\
&= \prod_{i=1}^n \{p^{\Delta_{i1}}(X_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = 1) \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)\} \\
&\times \frac{\prod_{i=1}^n \text{pr}(D_{i1} = 1 | S_i, \mathbf{Z}_{i1}) \prod_{j=2}^{M+1} \text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})}{\sum_{l=1}^{M+1} \text{pr}(D_{il} = 1 | S_i, \mathbf{Z}_{il}) \prod_{j \neq l}^{M+1} \text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})} \\
&= \prod_{i=1}^n \{p^{\Delta_{i1}}(X_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = 1) \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)\} \\
&\times \frac{\prod_{i=1}^n \text{pr}(D_{i1} = 1 | S_i, \mathbf{Z}_{i1}) / \text{pr}(D_{i1} = 0 | S_i, \mathbf{Z}_{i1})}{\sum_{l=1}^{M+1} \text{pr}(D_{il} = 1 | S_i, \mathbf{Z}_{il}) / \text{pr}(D_{il} = 0 | S_i, \mathbf{Z}_{il})}. \tag{2.9}
\end{aligned}$$

One may also note that there is an underlying asymmetry in the model in terms of treating the partially missing covariate X and completely observed covariate $\mathbf{Z}$ which I assume to be nonstochastic. One could model the completely observed covariate $\mathbf{Z}$ in a retrospective fashion as well by assuming a parametric distribution for $p(\mathbf{Z}_{ij} | D_{ij} = 0, S_i)$. Adopting that model will lead to

$$\text{pr}(D_{ij} = 1 | S_i) / \text{pr}(D_{ij} = 0 | S_i) = \int \frac{\text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i)}{\text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i)} \times p(\mathbf{Z}_{ij} | D_{ij} = 0, S_i) d\mathbf{Z}_{ij}.$$

This imposes assumptions on the unconditional disease prevalence in each stratum which is harder to justify in comparison to (2.5). More importantly, if that specified model for $p(\mathbf{Z}_{ij} | D_{ij} = 0, S_i)$ is incorrect, one risks the possibility of

biased, and inconsistent estimation. Since $\mathbf{Z}$ is usually multivariate with a mixture of categorical and continuous covariates, specifying an accurate model for its distribution may not be an easy task.

Using (2.5)-(2.7), the conditional likelihood is given by

$$\begin{aligned}
& L_c(\beta_1, \beta_2, \gamma, \gamma_{01}, \gamma_{02}, \dots, \gamma_{0n}) \\
& \propto \left\{ \prod_{i=1}^n p^{\Delta_{i1}}(X_{i1} | \mathbf{Z}_{i1}, S_i, D_{i1} = 1) \times \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij} | \mathbf{Z}_{ij}, S_i, D_{ij} = 0) \right\} \\
& \quad \times \frac{\exp[\sum_{i=1}^n \beta_1^T \mathbf{Z}_{i1} + \sum_{i=1}^n \xi_{i1} \{(b(\theta_{i1}^*) - b(\theta_{i1}))\}]}{\prod_{i=1}^n [\sum_{j=1}^{M+1} \exp[\beta_1^T \mathbf{Z}_{ij} + \xi_{ij} \{b(\theta_{ij}^*) - b(\theta_{ij})\}]]} \\
& = \prod_{i=1}^n \exp[\xi_{i1} \Delta_{i1} \{\theta_{i1}^* X_{i1} - b(\theta_{i1}^*)\} + \Delta_{i1} c(X_{i1}, \xi_{i1})] \\
& \quad \times \prod_{i=1}^n \prod_{j=2}^{M+1} \exp[\Delta_{ij} \xi_{ij} \{\theta_{ij} X_{ij} - b(\theta_{ij})\} + \Delta_{ij} c(X_{ij}, \xi_{ij})] \\
& \quad \times \prod_{i=1}^n \left(1 + \sum_{j=2}^{M+1} \exp \left[\beta_1^T (\mathbf{Z}_{ij} - \mathbf{Z}_{i1}) + \xi_{ij} \{b(\theta_{ij}^*) - b(\theta_{ij})\} \right. \right. \\
& \quad \left. \left. - \xi_{i1} \{b(\theta_{i1}^*) - b(\theta_{i1})\} \right] \right)^{-1}.
\end{aligned} \tag{2.10}$$

The main problem in (2.10) is to estimate the regression parameters β_1 and β_2 . However, it may be noted that while the above conditional likelihood has eliminated the nuisance parameters $\beta_0(S_i)$, the nuisance parameters γ_{0i} remain (via θ_{ij}), and they increase in direct proportion to the number of strata. Hence, even the conditional maximum likelihood of β_1 or β_2 is potentially subject to the Neyman-Scott phenomenon, and may produces inconsistent estimators. Adopting a Bayesian approach is one way to circumvent this difficulty.

I consider the following set of mutually independent priors, $\beta_1 \sim \text{Normal}(\mu_{\beta_1}, \Sigma_{\beta_1})$, $\beta_2 \sim \text{Normal}(\mu_{\beta_2}, \Sigma_{\beta_2}^2)$, $\gamma_j \stackrel{iid}{\sim} \text{Normal}(\mu_*, \sigma_*^2)$, and $\gamma_{0i} | G \stackrel{iid}{\sim} G$, where G has a Dirichlet process prior with a normal base measure. Symbolically, $G \sim DP(\alpha G_0)$

and G_0 is $N(\zeta_0, \sigma_0^2)$, where α is the concentration parameter. Using the result of Antoniak (1974), it follows that the predictive distribution of $\gamma_{0\overline{n+1}}$ given $\gamma_{01}, \dots, \gamma_{0n}$ is

$$\gamma_{0\overline{n+1}} | \gamma_{01}, \dots, \gamma_{0n} \sim \frac{\alpha}{\alpha + n} N(\cdot; \zeta_0, \sigma_0^2) + \frac{1}{\alpha + n} \sum_{k=1}^n I_{\gamma_{0k}}(\cdot). \quad (2.11)$$

Remark 1. Note that for very large values of α relative to n , the Bayesian semiparametric method produces essentially the $N(\zeta_0, \sigma_0^2)$ predictive distribution for $\gamma_{0\overline{n+1}}$, while very small values of α relative to n amounts to the fact that the predictive distribution of $\gamma_{0\overline{n+1}}$ essentially equals the “empirical” distribution of $\gamma_{01}, \dots, \gamma_{0n}$.

One may also note that as $\alpha \rightarrow \infty$ the Dirichlet process model reduces to specifying a parametric model on the γ_{0i} , namely $\gamma_{0i} \stackrel{iid}{\sim} G_0$, whereas $\alpha = 0$ simply implies a parametric model with a constant stratum effect, namely, $\gamma_{0i} = \gamma_0$ and $\gamma_0 \sim G_0$.

In my computations, I assume a Gamma prior on α and resample from the full conditional distribution of α using a latent beta variable as described in Escobar and West (1995).

Let

$$g_{ij} = \exp[\beta_1^T (\mathbf{Z}_{ij} - \mathbf{Z}_{i1}) + \xi_{ij} \{b(\theta_{ij}^*) - b(\theta_{ij})\} - \xi_{i1} \{b(\theta_{i1}^*) - b(\theta_{i1})\}]. \quad (2.12)$$

Now the full conditional distribution of the parameters may be obtained as:

$$\pi(\beta_1 | \cdot) \propto \prod_{i=1}^n (1 + \sum_{j=2}^{M+1} g_{ij})^{-1} \times \exp[-\frac{1}{2}(\beta_1 - \mu_{\beta_1})^T \Sigma_{\beta_1}^{-1}(\beta_1 - \mu_{\beta_1})]; \quad (2.13)$$

$$\pi(\beta_2 | \cdot) \propto \exp[\sum_{i=1}^n \Delta_{i1} \xi_{i1} \{\theta_{i1}^* X_{i1} - b(\theta_{i1}^*)\}] \times \prod_{i=1}^n (1 + \sum_{j=2}^{M+1} g_{ij})^{-1} \times \exp[-\frac{1}{2\sigma_{\beta_2}^2}(\beta_2 - \mu_{\beta_2})^2]; \quad (2.14)$$

$$\begin{aligned}
\pi(\gamma_j|\cdot) &\propto \exp\left[\sum_{i=1}^n \Delta_{i1}\xi_{i1}\{\theta_{i1}^*X_{i1} - b(\theta_{i1}^*)\}\right] \times \exp\left[\sum_{i=1}^n \sum_{j=2}^{M+1} \Delta_{ij}\xi_{ij}\{\theta_{ij}X_{ij} - b(\theta_{ij})\}\right] \\
&\times \prod_{i=1}^n \left(1 + \sum_{j=2}^{M+1} g_{ij}\right)^{-1} \times \exp\left[-\frac{1}{2\sigma_j^2}(\gamma_j - \mu_j)^2\right], \quad j = 1, \dots, p; \quad (2.15)
\end{aligned}$$

$$\begin{aligned}
\pi(\gamma_{0i}|\cdot) &\propto \frac{\alpha}{\alpha + n - 1} w_1(x_{ij}, z_{ij}, \beta_1, \beta_2, \gamma_1, \Delta_{ij}, \theta_{ij}, \xi_{ij}) \times N(\zeta_0 + \sigma_0^2 \sum_{j=1}^{M+1} \Delta_{ij}\xi_{ij}x_{ij}, \sigma_0^2) \\
&+ \frac{1}{\alpha + n - 1} w_2(x_{ij}, z_{ij}, \beta_1, \beta_2, \gamma_1, \Delta_{ij}, \theta_{ij}, \xi_{ij}) \sum_{\substack{k=1 \\ k \neq i}}^n I_{\gamma_{0k}}(\gamma_{0k}), \quad (2.16)
\end{aligned}$$

where I is the indicator function and w_1 and w_2 are two weight functions which can be calculated explicitly and are such that $\pi(\gamma_{0i}|\cdot)$ is a proper density. Thus expression (2.16) may be derived using the classical results of Antoniak (1974). We also note that θ_{ij} and θ_{ij}^* involve γ_j .

For estimation of parameters I use a Markov chain Monte Carlo computing scheme. The simulation strategy is discussed in detail in Section 2.6.

Remark 2. It is easy to extend results of this section to multiple exposures

$\mathbf{X}_{ij} = (X_{ij1}, \dots, X_{ijq})^T$ where the X_{ijk} are mutually independent, each having a distribution belonging to the exponential family. One then needs to introduce the indicator variables $\Delta_{ijk} = 1$ or 0 according as X_{ijk} is observed or missing. The likelihood given in (8) then needs to be suitably modified. The extension to the case when these multiple exposures may possibly be associated is more involved as one then need to model the association structure in a suitable way. These issues for multivariate exposures are considered in Chapter 5.

2.4 Computational Details

Note that none of the full conditionals will has a standard distributional form, and thus simulating observations from the conditionals is not automatic. I adopt a componentwise Metropolis-Hastings algorithm for each of the parameters, namely

β_1, β_2, γ . For simulating observations from the posterior of γ_{0i} I use an algorithm suggested by Neal (2000). The details of the computational steps are as follows.

For parameters other than γ_{0i} , $i = 1, \dots, n$ I essentially follow the same Metropolis-Hastings scheme. Let θ stand for the generic parameter, i.e., any of β_1, β_2, γ . Let $L_c(\theta|\cdot)$ denote the conditional likelihood as furnished in (2.10) as a function of θ given the data and all other parameters. Let $\pi(\theta)$ be the prior distribution on θ . In order to simulate observations from the full conditional distribution of θ , namely $\pi(\theta|\cdot)$ I follow the following steps.

Step 1. Start with any reasonable initial value of θ , say θ_0 . This is the current value for θ .

Step 2. Generate a new value of θ , say θ^* from a candidate density $g(\theta)$.

Step 3. Replace θ_0 by θ^* with probability $\rho(\theta, \theta^*) = \min\{1, \frac{\pi(\theta^*|\cdot)g(\theta_0)}{\pi(\theta_0|\cdot)g(\theta^*)}\}$.

Retain the existing value of θ_0 otherwise.

Remark 3. I choose the candidate density $g(\theta)$ to be identical to the prior density $\pi(\theta)$. Note that the full conditional density of θ , namely, $\pi(\theta|\cdot) \propto \pi(\theta)L_c(\theta|\cdot)$. Consequently, the acceptance probability as described in Step 2 above, reduces to (after cancelation of the prior term with the identical candidate density term),

$$\min\{1, \frac{L_c(\theta^*|\cdot)}{L_c(\theta_0|\cdot)}\}.$$

Once I update the value of the first set of parameters, say, β_1 , I move on to the similar cycle for β_2 with the updated value of β_1 substituted in the likelihood in (2.10). After updating β_2 , one can proceed to the Metropolis-Hastings step for γ_1 with the updated values of β_1 and β_2 , and finally with all three updated values of $\beta_1, \beta_2, \gamma_1$ to the iteration steps for γ_{0i} , $i = 1, \dots, n$.

Note that the convergence of this Markov chain to the proper conditional distribution $\pi(\theta|\cdot)$ is automatic because the transition kernel

$$k(\theta, \theta^*) = \rho(\theta, \theta^*)\pi(\theta^*) + (1 - r(\theta))\delta_\theta(\theta^*), \quad (2.17)$$

where $r(\theta) = \int (1 - \rho(\theta, \theta^*))\pi(\theta^*)d\theta^*$ and δ_θ denotes the Dirac mass at θ , satisfies the *detailed balance* condition (see Robert and Casella, 2001 for details) i.e.,

$$k(\theta, \theta^*)\pi(\theta|\cdot) = k(\theta^*, \theta)\pi(\theta^*|\cdot).$$

Remark 4. For the equine example in Section 2.5.1 with the exposure (age) distribution modeled normally there will be an additional scale parameter σ^2 on which I will assume an Inverse Gamma prior. In that case there is an additional Metropolis Hastings step for σ^2 similar to that described above.

Generating observations from conditional of γ_{0i} : For simulating observations from the full conditional of γ_{0i} , $i = 1, \dots, n$, first note that this will be substantially computation-intensive as n is typically large in many of the real examples. I adopt one of the algorithms suggested by Neal (2000) for generating observations from the posterior of Dirichlet mixtures.

To illustrate the computational steps, let us fix our attention to a particular stratum, say, stratum k . Also let $\gamma_{0(-k)} = (\gamma_{01}, \dots, \gamma_{0\ k-1}, \gamma_{0\ k+1}, \dots, \gamma_{0n})$, the vector of all but the k^{th} stratum effect parameter. I proceed in the following way:

Step 1. Start with some reasonable values of γ_{0i} , $i = 1, \dots, n$. The current state of the Markov chain consists of these n values, namely, $(\gamma_{01}, \dots, \gamma_{0k}, \dots, \gamma_{0n})$.

Step 2. Draw a candidate value γ_{0k}^* from the distribution of $\gamma_{0k}|\gamma_{0(-k)}$, namely, from,

$$(\alpha + n - 1)^{-1} \sum_{j \neq k}^n I_{\gamma_{0j}}(\cdot) + \{\alpha/(\alpha + n - 1)\}N(\cdot; \zeta_0, \sigma_0^2).$$

There are several ways of generating observations from the above mixture distribution, which essentially reflects that for the k -th stratum effect you either get

a new distinct value from the normal component of the mixture with probability $\alpha/(\alpha + n - 1)$, otherwise it is drawn with equal probability from the current set of $(n - 1)$ entries of $\gamma_{0(-k)}$, the vector corresponding to the other $(n - 1)$ stratum effect parameters. Following is one particular way to generate observations from this candidate density.

A) Let $p_j = 1/(\alpha + n - 1)$ for $j = 1, \dots, n$, $j \neq k$ and $p_k = \alpha/(\alpha + n - 1)$.

B) Calculate the cumulative probabilities $c_l = \sum_{j=1}^l p_j$. Let $c_0 \equiv 0$.

C) Draw a random number u from $U(0,1)$ distribution. Then $c_{j-1} < u \leq c_j$ for some specific $j \in \{1, \dots, n\}$. If $j = k$, draw the candidate γ_{0k}^* from $N(\zeta_0, \sigma_0^2)$. If $j \neq k$, take $\gamma_{0k}^* = \gamma_{0j}$.

Step 3. Compute the acceptance probability $a(\gamma_{0k}^*, \gamma_{0k}) = \min\{1, L_c(\gamma_{0k}^*|\cdot)/L_c(\gamma_{0k}|\cdot)\}$, where $L_c(\gamma_{0k}^*|\cdot)$, and $L_c(\gamma_{0k}|\cdot)$ are the conditional likelihoods as given in (2.10) evaluated at γ_{0k}^* and γ_{0k} , respectively, with all other parameters held fixed at their current values.

Step 4. Set the new value of γ_{0k} to γ_{0k}^* with the above probability.

Step 5. Repeat Steps 2-4 R times. I choose $R = 5$ in all my computations.

Consider the last of these R updates of γ_{0k} as the current value of the k -th stratum effect parameter. I repeat the above steps for all n stratum effect parameters, so for the equine data, there will be 498 such cycles to be performed, one for each of the γ_{0i} , $i = 1, \dots, n$.

Remark 5. For resampling from the full conditional distribution of α I follow the algorithm suggested by Escobar and West (1995). At each step I count the number of distinct γ_{0i} , say k , and conditional on the current values of α , and k , I simulate a latent beta random variable say, $\eta \sim B(\alpha + 1, n)$. Let $G(a, b)$ denote the prior on α . Using the current value of k , and η , α is simulated from the following mixture of Gamma distribution,

$$\pi(\alpha|\eta, k) = pG(a + k, b - \log(\eta)) + (1 - p)G(a + k - 1, b - \log(\eta))$$

where $p = (a + k - 1) / [a + k - 1 + n\{b - \log(\eta)\}]$.

One complete step of the entire Markov chain consists of the above componentwise Metropolis-Hastings steps for all the parameters in the model. I proceed iteratively through the whole cycle for all the parameters until convergence. I run the chain typically from 7000-10000 iterations, and calculate the diagnostic proposed by Gelman and Rubin (1992) as a measure of convergence. I furnish the posterior means as my estimates for the parameters, and provide as well the highest posterior density credible region for all the parameters. The codes for implementing the methodology are available on my homepage <http://www.stat.ufl.edu/~ssinha>.

2.5 Two Special Cases with Data Examples

It is worth noting that the proposed model can accommodate both continuous and discrete exposure variables as well as modeling for the missingness in the exposure variable. In this section, I present two particular illustrations, one with a normally distributed exposure variable and the other with a binary exposure variable, all motivated by real examples.

2.5.1 Continuous Exposure: Equine Epidemiology Example

The first example is the equine epidemiology example described in the introduction. Age was considered as a single exposure variable (X) measured on a continuous scale with one binary covariate (Z) indicating whether the horse experienced recent diet changes or not. For scaling purposes I linearly transformed age so that X was on the interval $[0, 1]$. Since I am dealing with a single covariate and one exposure variable here, the corresponding regression parameters will all be scalar.

The proposed method for the general exponential family, as described in the previous section, applies here with continuous exposure X . I assume that conditional on $D_{ij} = 0$, i.e., for the control population the exposure variable age has a normal distribution of the following form :

$$[X_{ij} | D_{ij} = 0, Z_{ij}, S_i] \sim \text{Normal}(\gamma_{0i} + \gamma_1 Z_{ij}, \sigma^2), \quad (2.18)$$

where γ_{0i} are stratum effects on the exposure distribution and γ_1 is the regression parameter for regressing X_{ij} on the measured covariates Z_{ij} . I also assume the logistic regression model for disease status (2.1). From (2.8) it follows the exposure distribution among the cases is Normal with mean $\gamma_{0i} + \gamma_1 Z_{ij} + \sigma^2 \beta_2$ and variance σ^2 . The conditional likelihood analogous to (2.10) can now be derived using these facts. I used the following prior distributions. The parameters $(\beta_1, \beta_2, \gamma_1)$ were independent Normal(0, 10), and the γ_{0i} 's had a Dirichlet prior with base measure $G_0 \sim \text{Normal}(0, 10)$. Experimentation suggested that the parameter estimates remain relatively unchanged over a range of Gamma priors on α . Table 2-1 contains the results corresponding to a Gamma(2, 2) prior on α . The variance σ^2 in (2.18) is assigned an inverse gamma prior, i.e. $\pi(\sigma^2) \propto (\sigma^2)^{-(c/2)-1} \exp(-\frac{d}{2\sigma^2})$, written symbolically as IG($c/2, d/2$): for my example $c = 60$ and $d = 2$.

The full conditional distributions for all the parameters have compact expressions that can be derived as special cases of equations (2.13)-(2.16). The full conditional distribution for the additional scale parameter σ^2 is given by,

$$\pi(\sigma^2 | \cdot) \propto \sqrt{\frac{\lambda}{2\pi}} \frac{1}{(\sigma^2)^{3/2}} \exp\left[-\frac{\lambda}{2\mu^2} \frac{(\sigma^2 - \mu)^2}{\sigma^2}\right] (\sigma^2)^{-\frac{2n+c-1}{2}}, \quad (2.19)$$

where $\lambda = -2(a - \frac{d}{2})$, $\mu = \sqrt{\frac{(a-d)}{b}}$, $a = -\frac{1}{2} \sum_{i=1}^n \{(X_{i1} - \theta_{i1})^2 + (X_{i2} - \theta_{i2})^2\}$, $b = -\frac{n}{2} \beta_2^2$.

For estimation of the parameters I adopt a Markov chain Monte Carlo numerical integration scheme. For simulating observations from the posterior distribution of γ_{0i} I follow the idea of Escobar (1994), West, Müller and Escobar (1994), Escobar and West (1995), and Neal (2000). For generating random numbers from the conditional distribution of other parameters I follow a standard componentwise Metropolis-Hastings scheme (Robert and Casella, 2001). The details of the computing scheme are provided in Section 2.4. For each of my examples and simulated datasets, I conduct three analyses. One is my proposed Bayesian semiparametric

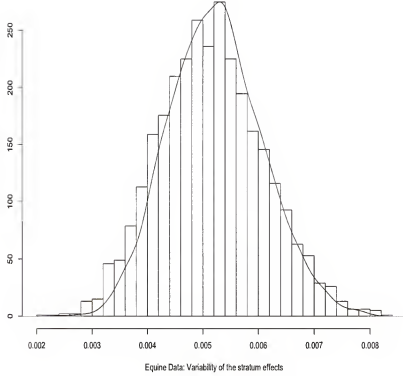


Figure 2-1: Plot of the variability of γ_{0i} 's for the equine data example. Estimates of 498 γ_{0i} 's were collected for each of the last 5000 MCMC samples. Variance of these 498 values were then calculated for each run. The histogram is of these 3000 variance values with a kernel density estimate overlayed on it.

analysis (BSP). The other is a parametric Bayesian (PB) version of the work of Satten and Carroll (2000) by considering a constant stratum effect $\gamma_{0i} \equiv \gamma_0$ and then assuming a normal prior on this common stratum effect parameter γ_0 . The parametric Bayesian method is simply the Bayesian analog of the method of Satten and Carroll (2000). Also, note that the PB method is a special case of BSP method with α equal to zero.

I also present the frequentist alternative for analyzing this matched data through a standard conditional logistic regression (CLR) analysis (Breslow and Day, 1980).

I ran the three analyses mentioned above on the equine dataset with all 498 strata. The three analysis were reran where I randomly deleted 40% of the age observations. The results are given in Table 2-1. Briefly, with no missing data the

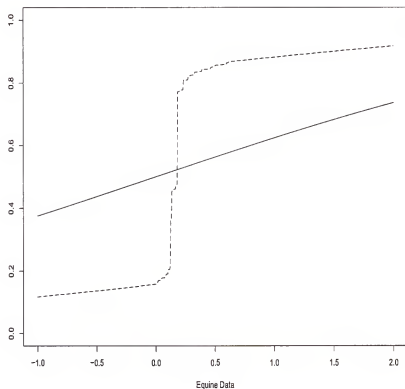


Figure 2-2: Predictive distribution of γ_{0n+1} for the equine data example. The dashed line shows the predictive distribution of γ_{0n+1} for the Equine data example overlayed with the solid line for the distribution function of the base measure $N(0, 10)$.

results of the methods are roughly in accordance with one another. The slightly small standard error estimates for the Bayesian methods may be a reflection of the

fact that unlike CLR, they are full likelihood based semiparametric and parametric methods. With missing data, of course, one sees that the Bayesian methods yield much smaller standard errors than CLR, reflecting the fact that proposed methods use all the data, and not just the pairs with no missing data. There is little difference between the semiparametric analysis and the parametric Bayes version of the method of Satten and Carroll (2000), a phenomenon that can be explained as follows. Figure 2-1 shows a density plot of the variance of γ_{0i} 's across the 498 strata in the last 5000 MCMC samples. The average variance among the γ_{0i} 's is very small (about 0.0054), suggesting that the stratum effects are almost constant across strata. Thus the parametric model with constant stratum effect essentially holds for this example, and hence it is not surprising that the two methods give similar results. Figure 2-2 presents a plot for the predictive distribution of γ_{0n+1} given the data and it's evident from this figure that the distribution is different from simple normal model.

2.5.2 Binary Exposure: Endometrial Cancer

In the Los Angeles 1:4 matched case-control study on endometrial cancer as discussed in Breslow and Day (1980), the cases were matched with four controls on the basis of age, the community the case lived in, marital status, and the time the case had entered in the community. The data has measurement on several binary covariates: presence of gall bladder disease, hypertension, use of estrogen, to name a few. The variable obesity may be considered as a binary exposure variable (X) for contracting endometrial cancer. This variable had about 16% missing observations.

I considered the presence of Gall Bladder disease in the subject as the completely observed dichotomous covariate Z .

In such a case with dichotomous exposure variable, the exposure distribution for control population is naturally assumed to have a logistic form: $\text{pr}(X_{ij} =$

$1|Z_{ij}) = H(\gamma_{0i} + \gamma_1 Z_{ij})$, where $H(\cdot)$ is the logistic distribution function. From (2.8) the exposure distribution in the case population is again of the logistic form with: $\text{pr}(X_{ij} = 1|Z_{ij}) = H(\gamma_{0i} + \gamma_1 Z_{ij} + \beta_2)$.

In the LA study there were 63 strata. For this example also I adopted the same Metropolis Hastings scheme by Neal (2000) for simulating observations from the posterior of γ_{0i} 's. Alternately, one can use the approach of MacEachern and Müller (1998), but Neal's approach seems easier to implement in my case. A detailed outline of the algorithm and computing scheme is already given in Section 2.4.

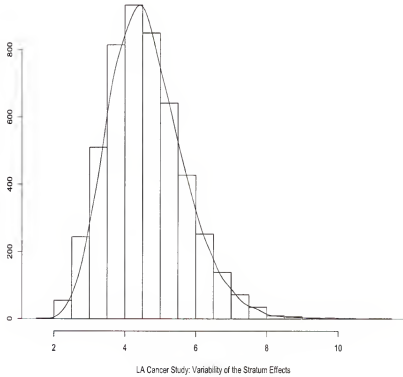


Figure 2-3: Plot of the variability of γ_{0i} 's for the LA cancer study example. Estimates of the 63 γ_{0i} 's were collected for each of the last 5000 MCMC samples. Variance of these 63 values were then calculated for each run. The histogram is of these 3000 variance values with a kernel density estimate overlaid on it.

I present the results with a Gamma (2, 2) prior on α . I considered a Normal(0, 10) prior on all the parameters as well as the base measure in the Dirichlet process prior.

The original data contained missing observations on the exposure variable obesity. I modeled the missingness in the exposure variable in both the parametric Bayes (the Bayesian version of the Satten and Carroll method with fixed stratum effects) and the semiparametric Bayes analysis. Classical conditional logistic analysis is also conducted. Table 2-2 presents the results. Although there are certainly some interesting numerical differences, in the main the results are reasonably comparable. Figure 2-3 presents a density plot of the variance of the 63 stratum effects in the last 5000 MCMC samples with the BSP method. The average variance among the γ_{0i} 's is approximately 4.5. This is fairly large, and if there were major covariate effects one would expect the BSP and the parametric Bayes version of the method of Satten and Carroll (2000) to be different. However, in this example the analyses suggest that $\beta_2 = 0$ is plausible, so that D and X are essentially independent and whether the γ_{0i} 's are constant or not does not cause a model bias. This is an example where there is stratum variability and natural missingness in the data. Figure 2-4 presents a plot of the predictive distribution of γ_{0n+1} .

2.6 Simulation Study

In order to simulate a realistic dataset for comparing the Bayesian semiparametric method with the parametric Bayes version of the method of Satten and Carroll (2000) and standard matched conditional logistic regression, I used the equine data itself as a prototype, except that I reversed the roles of diet change and age. I generated a hypothetical 1:1 matched dataset with 50 strata and with now diet change (a binary variable) as the exposure (X) and age of horse (transformed on to [0,1]) as an observed covariate (Z). To elicit a realistic value for the true

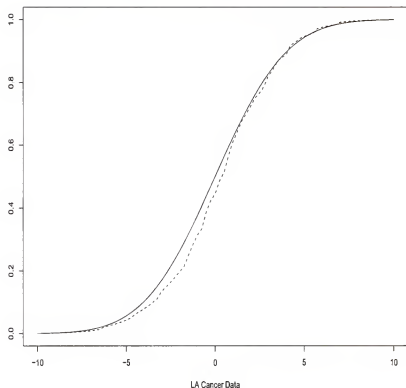


Figure 2-4: Predictive distribution of γ_{0n+1} for the LA cancer study example. The dashed line shows the predictive distribution of γ_{0n+1} for the Los Angeles Cancer Study example, overlaid with the solid line for the distribution function of the base measure $N(0, 10)$.

parameters $(\beta_1, \beta_2, \gamma_1)$ I made use of the original equine dataset at hand. I first performed a conditional logistic regression analysis of the full equine data with X and Z as covariates. The parameter estimates for the coefficients of X and Z using the CLR analysis were chosen as the true values of β_1 and β_2 in my simulation: these were 2.049, and 2.129, respectively. In order to elicit values for γ_1 , I ran a logistic regression with diet change X as the response and age Z as the covariate. The fitted equation had logits $-3.66 + 0.63Z$. In the simulation study I thus used $\gamma_1 = 0.63$. The purpose of the simulation was to explore the relative performance

of the three methods when in fact $\gamma_{0i} \equiv \gamma_0$, the standard model assumptions were true, and also for the situation when the constant stratum effect assumption does not hold and γ_{0i} came from a distribution with reasonable variability.

I followed the structure of my models as described in Section 2 for simulating X , D , and Z . I started first with generating the covariate Z . For the equine data, the histogram of the variable horse age resembled a distribution which could well be described by a Gamma distribution with shape and scale parameters 1.8 and 0.2 respectively. Accordingly, I generated covariate Z from the same Gamma distribution and then scaled the generated values to $[0,1]$. The results remain similar if one generates Z from a normal distribution.

Second, I generated the binary variable D standing for disease status. For the i -th stratum, it may easily be noted that,

$$Pr(D_{i1} = 1 | D_{i1} + D_{i2} = 1, Z_{i1}, Z_{i2}, S_i) = 1 - \frac{Q_{i2}}{Q_{i2} + \exp[\beta_1(Z_{i1} - Z_{i2})]Q_{i1}},$$

where, for $j = 1, 2$,

$$Q_{ij} = \frac{1 + \exp(\gamma_{0i} + \gamma_1 Z_{ij} + \beta_2)}{1 + \exp(\gamma_{0i} + \gamma_1 Z_{ij})}.$$

A Bernoulli variable was generated with the above specified probability structure and conditional on the fact that the simulated value for D_{i1} is a 1, i.e., the subject is a member of the case population, I generated the binary exposure mimicking diet change (X) from a Bernoulli distribution with $\text{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i1} + \beta_2$; if the simulated value for D_{i1} is 0, i.e., the subject is a member of the control population, I generate X from a Bernoulli distribution with $\text{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i1}$. Once we have simulated D_{i1} in the above manner, for the other observation in the i -th stratum, $D_{i2} = 1 - D_{i1}$. The corresponding exposure value X_{i2} is then simulated from a Bernoulli distribution with either $\text{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i2} + \beta_2$ (when $D_{i2} = 0$) or $\text{logit}(p_i) = \gamma_{0i} + \gamma_1 Z_{i2} + \beta_2$ (when $D_{i2} = 1$).

I performed two sets of simulations, one with a constant value of γ_{0i} namely, -3.66 (this choice is elicited above) the other with γ_{0i} simulated from a normal distribution with mean -3.66 and standard deviation 2. The results are presented in Table 2-4.

In this simulations I used identical parameters for the normal distribution with mean zero and variance 10, which was assumed to be the prior on γ_0 in the parametric Bayes version of the method of Satten and Carroll (2000) and also as the base measure in the Dirichlet Process prior for the semiparametric Bayesian method. I used a Gamma (2, 2) prior on α . I replicated the simulations 50 times, generating 50 different datasets, and obtained the parameter estimates by the three methods, and computed the mean and mean squared error of these 50 estimates. For each replication I also generated a dataset with 30% observations missing completely at random and recalculated all the estimates.

The results are fairly clear. In the case of fixed strata effects and no missing data, there are little in the way of differences among the Bayesian semiparametric method, the parametric Bayesian counterpart to the frequentist method of Satten and Carroll, and ordinary conditional logistic regression (CLR). However, if there is 30% missing data, then ordinary CLR is clearly less efficient. This might be looked upon as a quantification of robustness of efficiency for the Bayesian semiparametric method. On the other hand, if there are major unexplained stratum effects, i.e., the γ_{0i} have considerable variability, then one can see that the methods separate, with the Bayesian semiparametric method clearly dominating in terms of mean squared error.

2.7 Concluding Remarks

This Chapter considers matched case-control studies with missing exposure data. Parametric analyses require handling stratum effects on the distribution of

the exposure distribution. Previous methods either ignore such stratum effects, or assume that they can be captured in their entirety by a parametric model.

I have presented a Bayesian semiparametric approach to model the stratum effects. The methods proposed can handle discrete as well as continuous exposure, and also account for possible missingness in the exposure variable. The methodology is illustrated through real examples. Results based on simulated data indicate that in the presence of varying stratum effects, the semiparametric Bayesian method outperforms the parametric Bayesian method and frequentist methods currently in the literature.

Table 2-1: Results of the equine data example

		Full Equine Data			
Method		β_1	β_2	$10 \times \gamma_1$	$10 \times \sigma^2$
Bayes Semiparametric					
	Mean	2.16	2.18	0.18	0.23
	s.e.	0.33	0.39	0.14	0.01
	Lower HPD	1.57	1.36	-0.10	0.21
	Upper HPD	2.80	2.88	0.40	0.25
Bayes Parametric					
	Mean	2.14	2.10	0.22	0.26
	s.e.	0.32	0.41	0.17	0.01
	Lower HPD	1.60	1.32	-0.20	0.24
	Upper HPD	2.87	2.88	0.60	0.28
CLR					
	Mean	2.13	2.05		
	s.e.	0.32	0.47		
	Lower CL	1.50	1.13		
	Upper CL	2.76	2.97		
		Equine Data With 40% Missing Data			
Method		β_1	β_2	$10 \times \gamma_1$	$10 \times \sigma^2$
Bayes Semiparametric					
	Mean	2.16	2.30	0.14	0.23
	s.e.	0.32	0.48	0.24	0.01
	Lower HPD	1.70	1.30	-0.25	0.19
	Upper HPD	2.80	3.42	0.60	0.25
Bayes Parametric					
	Mean	2.13	1.90	0.17	0.28
	s.e.	0.32	0.47	0.20	0.02
	Lower HPD	1.56	0.92	-0.20	0.24
	Upper HPD	2.85	2.85	0.67	0.31
CLR					
	Mean	2.81	2.79		
	s.e.	0.59	0.77		
	Lower CL	1.65	1.47		
	Upper CL	3.96	4.49		

Here “Mean” is the posterior mean, “s.e.” is the posterior standard deviation, “Lower HPD” and “Upper HPD” are the lower and upper end of the HPD region respectively, “Lower CL” and “Upper CL” are the lower and upper end of the confidence limit respectively, The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

Table 2-2: Results of the endometrial cancer data example

		Gall Bladder	Obesity	
Method		β_1	β_2	γ_1
Bayes Semiparametric	Mean	1.29	0.70	-0.15
	s.e.	0.39	0.41	0.49
	Lower HPD	0.54	-0.10	-0.99
	Upper HPD	2.18	1.49	0.99
Bayes Parametric	Mean	1.19	0.49	0.42
	s.e.	0.39	0.36	0.62
	Lower HPD	0.58	-0.15	-0.70
	Upper HPD	2.12	1.25	1.78
CLR	Mean	1.28	0.44	
	s.e.	0.39	0.38	
	Lower CL	0.52	-0.30	
	Upper CL	2.04	1.19	

Here “Mean” is the posterior mean, “s.e.” is the posterior standard deviation, “Lower HPD” and “Upper HPD” are the lower and upper end of the HPD region respectively, “Lower CL” and “Upper CL” are the lower and upper end of the confidence limit respectively, The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

Table 2-3: Results of the simulation study for full data

Full data, fixed $\gamma_{0i} \equiv -3.66$				
Method		β_1	β_2	γ_1
Bayes Semiparametric	Mean	1.9880	2.0949	0.6938
	MSE	0.7595	0.5324	1.0342
Bayes Parametric	Mean	2.0089	1.8183	0.6743
	MSE	0.7785	0.5159	0.6454
CLR	Mean	2.3157	1.9496	
	MSE	2.0505	0.4603	
Full data, varying $\gamma_{0i} = \text{Normal}(-3.66, 4)$				
Method		β_1	β_2	γ_1
Bayes Semiparametric	Mean	1.7838	2.2194	0.4729
	MSE	0.7126	0.5042	0.8015
Bayes Parametric	Mean	1.7812	1.4641	0.4795
	MSE	0.9550	0.6488	0.8333
CLR	Mean	2.6350	1.7747	
	MSE	2.2622	0.5667	

Here “Mean” is the mean of the 50 estimates corresponding to the 50 simulated datasets while MSE is the mean squared error. The true parameter values are $\beta_1 = 2.049$, $\beta_2 = 2.129$, and $\gamma_1 = 0.63$. The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

Table 2-4: Results of the simulation study for missing data

30% missing data, fixed $\gamma_{0i} \equiv -3.66$				
Method		β_1	β_2	γ_1
Bayes Semiparametric	Mean	2.0022	2.4606	0.9617
	MSE	0.7925	0.8962	1.0962
Bayes Parametric	Mean	2.0141	1.8156	0.6913
	MSE	0.7905	0.8561	0.9125
CLR	Mean	2.9713	1.2689	
	MSE	5.6249	1.3182	
30% missing data, varying $\gamma_{0i} = \text{Normal}(-3.66, 4)$				
Method		β_1	β_2	γ_1
Bayes Semiparametric	Mean	1.7922	2.0150	0.4935
	MSE	0.8095	0.8152	1.0444
Bayes Parametric	Mean	1.7535	1.5081	0.5737
	MSE	0.9690	0.9444	1.3068
CLR	Mean	3.6817	1.2078	
	MSE	10.2676	1.3744	

Here “Mean” is the mean of the 50 estimates corresponding to the 50 simulated datasets while MSE is the mean squared error. The true parameter values are $\beta_1 = 2.049$, $\beta_2 = 2.129$, and $\gamma_1 = 0.63$. The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

CHAPTER 3

BAYESIAN SEMIPARAMETRIC MODELLING FOR MATCHED CASE-CONTROL STUDIES WITH MULTIPLE DISEASE STATES

3.1 Introduction

This Chapter considers matched case-control studies with multiple disease states. In some situations it is natural to note that the disease state might have more than one category, i.e., one may have subdivisions within the “cases”. For example, for patients diagnosed with cancer, they may have cancer of stage-I, stage-II or stage-III at the time of the diagnosis which is an example of ordinal disease categories. One may also notice nominal categories for disease states such as patients classified into having one eye or both eyes damaged. To the best of my knowledge there has been hardly any Bayesian, and very little frequentist work for analyzing matched data when one has multiple disease states. Along with considering polytomous disease states I extend the typical methodology for analyzing case-control data in the following way. To deal with partially missing exposure variable in this setup, I work with full conditional likelihood by assuming a prospective model of the disease status and a parametric distribution of the missing exposure variable in the control population given some observed covariates and stratum effect. More specifically, I try to capture unmeasured stratum heterogeneity through the distribution of the missing exposure variable. As a result, mentioned in Chapter 2, even the conditional likelihood involves a set of nuisance parameters due which grow in direct proportion to sample size. Therefore to handle the potential problem of inconsistency of the parameter estimates I took a semiparametric Bayesian approach by using a Dirichlet process prior on the stratum heterogeneity parameters.

The example considered in this Chapter involves a matched case-control dataset coming from a low-birth-weight study conducted by the Baystate Medical Center in Springfield, Massachusetts. The dataset is discussed in Hosmer and Lemeshow (2000, Section 1.6.2), and is used as an illustrative example of analyzing a matched case-control study in Chapter 7 of their book. Low-birth-weight, defined as birth weight less than 2500 grams, is a cause of concern for a newborn as infant mortality, and birth defect rates are very high for low-birth-weight babies. The data was matched according to the age of the mother. A woman's behavior during pregnancy (smoking habits, diet, prenatal care) can greatly alter the chances of carrying the baby to terms. The goal of the study was to determine whether these variables were "risk factors" in the clinical population served by Baystate Medical Center. In particular using the actual birth weight observations I divided the cases, namely, the low-birth-weight babies into two categories, *very low* (weighing less than 2000 gms), and *low* (weighing between 2000 to 2500 gms); and tried to assess the impact of smoking habits of mother on the chance of falling in the two low-birth-weight categories. Presence of uterine irritability in mother, and mother's weight at last menstruation period were considered as relevant covariates. It was noted that smoking mothers had a higher relative risk of having a low-birth-weight child when compared to a nonsmoking mother. However, the risk of having a *very low*-birth-weight child did not depend on smoking significantly. This observation could not be made without the classification of the data into multiple low-birth-weight groups, illustrating the relevance of such a type of analysis in certain situations.

The present Chapter intends to develop an approach to case-control studies with a multicategory variable D denoting disease states, a completely observed covariate Z , a vector of exposure variables or risk factors X which could potentially contain some missing observations. The exposure variables could be discrete or

continuous, and I assume that $\mathbf{X}|\mathbf{D}, \mathbf{Z}$ has a distribution coming from an exponential family which may have different natural parameters across strata. I assume a Dirichlet process with a normal base measure on the distribution of the stratum effects and normal priors on the other regression parameters, and estimate all the parameters via a Markov chain Monte Carlo computing scheme. This is a major departure from the usual frequentist as well as the Bayesian approach of assuming that the distribution of exposure variable is not affected by any stratum effect except through the measured covariates. This last assumption may not hold in many situations. For example, matching for cancer patients is often done from their family, and smoking is a natural exposure to consider. The distribution of the exposure may depend on genetic traits in the family which may affect the disease distribution in different families in different ways, and may not be measurable as a covariate. The present Bayesian approach will allow us to model stratum effects on the exposure distribution for highly stratified data, and account for missing exposure information.

The outline of the remaining sections is as follows. In Section 2, I introduce notations and model assumptions. Section 3 contains the conditional likelihood, the priors and the posterior distributions. In Section 4, I analyze the low-birth-weight data, and discuss the MCMC computation scheme. Section 5 presents the results from a small simulation study comparing the proposed Bayesian semiparametric method with two possible parametric Bayesian alternatives. Section 6 contains concluding remarks.

3.2 Model and Notation

Let D_{ij} denote the disease state of the j -th individual in the i -th stratum S_i . Suppose that there are $(K+1)$ nominal levels of the disease variable, with $D_{ij} = k$ denoting disease state k , $k = 1, \dots, K$, and $D_{ij} = 0$ denoting the control group. I assume that there is 1 case matched with M controls in each stratum,

and I have n strata in all. For ease of notation, I present my models with a single exposure X , but the model could easily be extended to the case when one has a set of independent multiple exposures each having a distribution coming from the exponential family. The disease probabilities for each of the K groups are modeled through K logits as in a multinomial logistic regression model:

$$\log \frac{\text{pr}[D_{ij} = k | S_i, \mathbf{Z}_{ij}, X_{ij}]}{\text{pr}[D_{ij} = 0 | S_i, \mathbf{Z}_{ij}, X_{ij}]} = \beta_{0k}(S_i) + \beta_{1k}^T \mathbf{Z}_{ij} + \beta_{2k} X_{ij} \quad \text{for } k = 1, \dots, K. \quad (3.1)$$

Each β_{1k} is a $p \times 1$ column vector, and $\mathbf{Z}_{ij} = (Z_{ij}^{(1)}, \dots, Z_{ij}^{(p)})^T$ is the vector of p completely observed covariates.

For the example with the low-birth-weight data as described in Section 3.1, I have two disease states with $K = 2$. Then $\exp(\beta_{21})$ signifies the odds of having low-birth-weight baby for a mother who smokes relative to one who does not smoke, and similarly $\exp(\beta_{22})$ is the risk of a *very* low-birth-weight child for a mother who smokes relative to one who does not. The conditional probabilities of the disease variable given the covariate, exposure, and the stratum are given by,

$$\text{pr}(D_{ij} = k | S_i, \mathbf{Z}_{ij}, X_{ij}) = \frac{\exp\{\beta_{0k}(S_i) + \beta_{1k}^T \mathbf{Z}_{ij} + \beta_{2k} X_{ij}\}}{1 + \sum_{r=1}^K \exp\{\beta_{0r}(S_i) + \beta_{1r}^T \mathbf{Z}_{ij} + \beta_{2r} X_{ij}\}} \quad \text{for } k = 1, \dots, K. \quad (3.2)$$

and

$$\text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij}, X_{ij}) = \frac{1}{1 + \sum_{r=1}^K \exp\{\beta_{0r}(S_i) + \beta_{1r}^T \mathbf{Z}_{ij} + \beta_{2r} X_{ij}\}}. \quad (3.3)$$

To cover both discrete as well as continuous exposures, we assume a general exponential family of distributions for the exposure variable in the control population with respect to some finite dominating measure μ , i.e.,

$$f(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) = \exp[\xi_{ij}\{\theta_{ij} X_{ij} - b(\theta_{ij})\} + c(\xi_{ij}, X_{ij})]. \quad (3.4)$$

The natural parameters are modeled as a regression function of the completely observed covariates, namely, $\theta_{ij} = \gamma_{0i} + \gamma_1^T \mathbf{Z}_{ij}$, where $\gamma_1^T = (\gamma_{11}, \dots, \gamma_{1p})$. The dependence of the exposure distribution on the stratum is captured through the varying intercepts γ_{0i} .

3.3 Likelihood, Priors and Posteriors

Before writing out the conditional likelihood one needs some technical results stated in the following as Lemmas. These Lemmas follow by repeating essentially the proofs of Lemmas 3 and 4 of Chapter 2. The details are omitted.

Lemma 1. Under assumptions (3.2)-(3.4) the distribution of the exposure variable in a given disease state k , namely, $f(X_{ij}|S_i, \mathbf{Z}_{ij}, D_{ij} = k)$ is also of general exponential form with scale parameter ξ_{ij} , and natural parameter $\theta_{ijk}^* = \theta_{ij} + \xi_{ij}^{-1} \beta_{2k}$ for $k = 1, \dots, K$.

Lemma 2. Under the same set of assumptions,

$$\frac{\text{pr}(D_{ij} = k|S_i, \mathbf{Z}_{ij})}{\text{pr}(D_{ij} = 0|S_i, \mathbf{Z}_{ij})} = \exp\{\beta_{0k}(S_i) + \beta_{1k}^T \mathbf{Z}_{ij}\} \times \exp[\xi_{ij}\{b(\theta_{ijk}^*) - b(\theta_{ij})\}]. \quad (3.5)$$

I need one more Lemma to write out the conditional likelihood.

Lemma 3.

$$\begin{aligned} & \frac{\text{pr}(D_{is} = k|S_i, \mathbf{Z}_{is})/\text{pr}(D_{is} = 0|S_i, \mathbf{Z}_{is})}{\sum_{j=1}^{M+1} \text{pr}(D_{ij} = k|S_i, \mathbf{Z}_{ij})/\text{pr}(D_{ij} = 0|S_i, \mathbf{Z}_{ij})} \\ &= \frac{\exp(\beta_{1k}^T \mathbf{Z}_{is}) \exp[\xi_{is}\{b(\theta_{isk}^*) - b(\theta_{is})\}]}{\sum_{j=1}^{M+1} \exp(\beta_{1k}^T \mathbf{Z}_{ij}) \exp[\xi_{is}\{b(\theta_{ijk}^*) - b(\theta_{ij})\}]} \text{ for } s = 1, \dots, M+1, i = 1, \dots, n. \end{aligned} \quad (3.6)$$

Without loss of generality one may assume that the first subject in each stratum is a case, and if one denotes the disease state (type of case) in stratum i as k_i (k_i could assume any of the values $1, \dots, K$), then the conditional likelihood given

that there is one case in each stratum will be of the form:

$$\begin{aligned}
L_c &\propto \prod_{i=1}^n \text{pr}\{D_{i1} = k_i, D_{ij} = 0 (j = 2, \dots, M+1), X_{ij} (j = 1, \dots, M+1) | S_i, \\
&\quad \mathbf{Z}_{ij}, \sum_{j=1}^{M+1} D_{ij} = k_i\} \\
&= \prod_{i=1}^n \{f(X_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = k_i) \prod_{j=2}^{M+1} f(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)\} \\
&\times \prod_{i=1}^n \frac{\text{pr}(D_{i1} = k_i | S_i, \mathbf{Z}_{i1}) \prod_{j=2}^{M+1} \text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})}{\sum_{l=1}^{M+1} \text{pr}(D_{il} = k_i | S_i, \mathbf{Z}_{il}) \prod_{j \neq l}^{M+1} \text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})} \\
&= \prod_{i=1}^n \{f(X_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = k_i) \prod_{j=2}^{M+1} f(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)\} \\
&\times \prod_{i=1}^n \frac{\text{pr}(D_{i1} = k_i | S_i, \mathbf{Z}_{i1}) / \text{pr}(D_{i1} = 0 | S_i, \mathbf{Z}_{i1})}{\sum_{j=1}^{M+1} \text{pr}(D_{ij} = k_i | S_i, \mathbf{Z}_{ij}) / \text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})}. \tag{3.7}
\end{aligned}$$

In case of partially missing exposure variable one should modify the above likelihood in the following way. Let δ_{ij} be an indicator variable for missing exposures defined in the following manner:

$$\delta_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{ is observed,} \\ 0 & \text{if } X_{ij} \text{ is missing, } i = 1, \dots, n \text{ and } j = 1, \dots, M+1. \end{cases}$$

I also assume that the distribution of δ_{ij} does not involve the parameters $\beta_{1k}, \beta_{2k}, \gamma_1$, and $\gamma_0^T = (\gamma_{01}, \dots, \gamma_{0n})$. The conditional likelihood including missingness of the exposure variable is seen to be,

$$\begin{aligned}
L_c &\propto \prod_{i=1}^n \{f(X_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = k_i)^{\delta_{i1}} \prod_{j=2}^{M+1} f(X_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)^{\delta_{ij}}\} \\
&\times \prod_{i=1}^n \frac{\text{pr}(D_{i1} = k_i | S_i, \mathbf{Z}_{i1}) / \text{pr}(D_{i1} = 0 | S_i, \mathbf{Z}_{i1})}{\sum_{j=1}^{M+1} \text{pr}(D_{ij} = k_i | S_i, \mathbf{Z}_{ij}) / \text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})}. \tag{3.8}
\end{aligned}$$

This likelihood involves the parameters $\beta_1^{p \times K} = (\beta_{11}, \dots, \beta_{1K}), \beta_2^T = (\beta_{21}, \dots, \beta_{2K}), \gamma_1$, and γ_0 .

I consider mutually independent normal priors: $\beta_{1k} \sim N_p(\mu_{\beta_{1k}}, \sigma_{\beta_{1k}}^2 \mathbf{I}_p)$ for $k = 1, \dots, K$, $\beta_2 \sim N_K(\mu_{\beta_2}, \sigma_{\beta_2}^2 \mathbf{I}_K)$, and $\gamma_1 \sim N_p(\mu_{\gamma_1}, \sigma_{\gamma_1}^2 \mathbf{I}_p)$. Following are the full conditional distributions of the parameters.

$$\pi(\beta_{1k}|\cdot) \propto \frac{\exp(-\frac{1}{2\sigma_{\beta_{1k}}^2} \mathbf{U}_{\beta_{1k}}^T \mathbf{U}_{\beta_{1k}})}{\prod_{i=1}^n \{\sum_{j=1}^{M+1} \exp(\mathbf{h}_i^T \beta_{11} Z_{ij}^{(1)} + \mathbf{h}_i^T \beta_{12} Z_{ij}^{(2)}) \times \frac{(1+\exp(\theta_{i1}^*))}{(1+\exp(\theta_{ij}))}\}\}, \text{ for } k = 1, 2, \quad (3.9)$$

$$\begin{aligned} \pi(\beta_2|\cdot) &\propto \frac{\exp(-\frac{1}{2\sigma_{\beta_2}^2} \mathbf{U}_{\beta_2}^T \mathbf{U}_{\beta_2})}{\prod_{i=1}^n \{\sum_{j=1}^{M+1} \exp(\mathbf{h}_i^T \beta_{11} Z_{ij}^{(1)} + \mathbf{h}_i^T \beta_{12} Z_{ij}^{(2)}) \times \frac{(1+\exp(\theta_{i1}^*))}{(1+\exp(\theta_{ij}))}\}\} \\ &\times \prod_{i=1}^n \{1 + \exp(\theta_{i1}^*)\}^{1-\delta_{i1}}, \end{aligned} \quad (3.10)$$

$$\begin{aligned} \pi(\gamma_1|\cdot) &\propto \frac{\exp(-\frac{1}{\sigma_{\gamma_1}^2} \mathbf{U}_{\gamma_1}^T \mathbf{U}_{\gamma_1})}{\prod_{i=1}^n \{\sum_{j=1}^{M+1} \exp(\mathbf{h}_i^T \beta_{11} Z_{ij}^{(1)} + \mathbf{h}_i^T \beta_{12} Z_{ij}^{(2)}) \times \frac{(1+\exp(\theta_{i1}^*))}{(1+\exp(\theta_{ij}))}\}\} \\ &\times \prod_{i=1}^n \left\{ \frac{(1 + \exp(\theta_{i1}^*))^{1-\delta_{i1}}}{(1 + \exp(\theta_{i1}))} \prod_{j=2}^{M+1} (1 + \exp(\theta_{ij}))^{-\delta_{ij}} \right\}, \end{aligned} \quad (3.11)$$

where

$$\begin{aligned} \mathbf{U}_{\beta_{1k}} &= \beta_{1k} - \mu_{\beta_{1k}} - \sigma_{\beta_{1k}}^2 \sum_{i=1}^n \mathbf{h}_i Z_{i1}^{(k)}, \\ \mathbf{U}_{\beta_2} &= \beta_2 - \mu_{\beta_2} - \sigma_{\beta_2}^2 \sum_{i=1}^n \delta_{i1} \mathbf{h}_i X_{i1}, \\ \mathbf{U}_{\gamma_1} &= \gamma_1 - \mu_{\gamma_1} - \sigma_{\gamma_1}^2 \sum_{i=1}^n \sum_{j=1}^{M+1} \delta_{ij} Z_{ij} X_{ij}, \end{aligned}$$

and $\mathbf{h}_i$ is defined as $\mathbf{h}_i = (h_{i1}, \dots, h_{iK})^T$ where,

$$h_{ir} = \begin{cases} 1 & \text{if } D_{i1} = r \\ 0 & \text{otherwise, } i = 1, \dots, n \text{ and } r = 1, \dots, K. \end{cases}$$

$$\begin{aligned} \pi(\gamma_{0i}|\cdot) &\propto \sum_{k \neq i} \frac{1}{\alpha + n - 1} w_k(\beta_{11}, \beta_{12}, \beta_2, \gamma_1, \gamma_{0s}, s \neq i) I(\gamma_{0k}) L_c \\ &+ w_i(\beta_{11}, \beta_{12}, \beta_2, \gamma_1, \gamma_{0s}, s \neq i) f_i(\gamma_{0i}|\cdot), \end{aligned} \quad (3.12)$$

where

$$f_i(\gamma_{0i}|\cdot) \propto \frac{\exp[-\frac{1}{2\sigma_0^2}(\gamma_{0i} - \xi_0 - \sigma_0^2 \sum_{i=1}^n \sum_{j=1}^{M+1} \delta_{ij} X_{ij})^2]}{\prod_{i=1}^n \{\sum_{j=1}^{M+1} \exp(h_i^T \beta_{11} Z_{ij}^{(1)} + h_i^T \beta_{12} Z_{ij}^{(2)}) \times \frac{(1+\exp(\theta_{i1}^*))}{(1+\exp(\theta_{ij}))}\} \\ \times \prod_{i=1}^n \{\frac{(1+\exp(\theta_{i1}^*))^{1-\delta_{i1}}}{(1+\exp(\theta_{i1}))} \prod_{j=2}^{M+1} (1+\exp(\theta_{ij}))^{-\delta_{ij}}\}$$

and $w_r, r = 1, \dots, n$ are weight functions such that $\pi(\gamma_{0i}|\cdot)$ is a proper density. One may note that if X is completely observed $\delta_{ij} = 1 \forall i, j$.

One of the key features of my approach is to retain the nuisance parameters γ_{0i} in the model, and to put a Dirichlet with normal mixture prior on them, i.e., $\gamma_{0i}|G \stackrel{iid}{\sim} G$, where $G \sim DP(\alpha G_0)$ and G_0 is $N(\zeta_0, \sigma_0^2)$. As mentioned in Chapter 2 and using Antoniak (1974), I have

$$\gamma_{0i}|\gamma_{0k}(k \neq i) \sim \frac{\alpha}{\alpha + n - 1} N(\cdot; \zeta_0, \sigma_0^2) + \frac{1}{\alpha + n - 1} \sum_{k=1, k \neq i}^n I_{\gamma_{0k}}(\cdot), \quad (3.13)$$

where I is the indicator function. The result allows the possibility of equal values of some γ_{0i} as well. As one will notice in my simulations, the fact that it can allow for equal values of γ_{0i} , plays an important role in making the procedure robust over a wide spectrum of scenarios, from widely varying γ_{0i} 's to the case when they are all equal.

For the data analysis, and simulations I compared the Bayesian semiparametric (BSP) method with two possible parametric Bayesian alternatives. The first one is a parametric Bayesian analogue of the method proposed by Satten and Carroll (2000) for matched case control studies with a single disease state. Satten and Carroll (2000) assumed a constant stratum effect on the exposure distribution, i.e., $\gamma_{0i} \equiv \gamma_0$. In the parametric Bayesian analogue of their method (denoted by PBC, C standing for constant stratum effect) in the context of multiple disease states, I consider a normal distribution as a prior on this common stratum effect parameter γ_0 , and carry out Bayesian analysis. The other parametric Bayesian

alternative (denoted by PBV, V standing for varying stratum effects) allows for possibly varying γ_{0i} , and assumes i.i.d. normal prior on each γ_{0i} .

Remark 1. It follows from (3.13) that for very large values of α the BSP method is equivalent to the PBV method whereas PBC is a special case of BSP method with α equals to zero. In my numerical work I assumed a Gamma prior on α , and resampled from the full conditional distribution of α using a latent beta variable as prescribed in Escobar and West (1995).

Remark 2. The entire analysis carries through if each stratum contains varying number of controls, with equations (3.5)-(3.9) remaining essentially the same with M replaced by M_i in the i -th stratum.

The estimation of the parameters is done by the Markov chain Monte Carlo numerical integration scheme. To generate random numbers from the posterior distributions of the parameters I use a componentwise Metropolis Hastings scheme. The computation scheme along with the analysis of the low birth-weight study are described in the following section.

3.4 Example and Computing Scheme

In the previous sections I discussed the general methodology which I now apply to the matched case-control study for low-birth-weight data as described in Section 1. The matched data contain 29 strata, and each stratum has one case, and 3 controls. I denote the low-birth-weight, and the *very* low-birth-weight group as disease state 1, and 2 respectively. One can possibly think of many different models for explaining the disease in terms of the possible covariates recorded in the data set. I consider smoking status of mother as a single exposure variable. Two other covariates, a binary variable denoting presence of uterine irritability (UI) in mother, and weight of the mother at last menstrual period (LWT) are also included in the model.

As a starting point, I separated the 29 strata into two groups depending on whether the case belonged to low-birth-weight category 1 or 2. I formed two cross-classification tables of birth weight category versus smoking status of mother during pregnancy period for these two separate matched samples, and noted that, $\widehat{\text{OR}}(1, 0) = 3.4$, and $\widehat{\text{OR}}(2, 0) = 1.917$, where $\widehat{\text{OR}}(k, 0)$ denotes the odds ratio of maternal smoking habits for birth weight group = k versus normal birth weight group = 0. The two odds ratios demonstrate that the odds of having a low-birth-weight baby for a smoking mother as opposed to a nonsmoking mother is higher in category 1, whereas for category 2, I notice a relatively weaker behavior in the same direction. The difference in the odds ratios in these two tables led us to use this data as a testbed example to illustrate my methods.

For the proposed analysis I have a stochastic distribution on the exposure variable which belongs to the exponential family. The binary variable smoking status is assumed to follow a Bernoulli distribution:

$$f(X_{ij}|D_{ij} = 0, \mathbf{Z}_{ij}, S_i) = p_{ij}^{X_{ij}} (1 - p_{ij})^{1-X_{ij}}, \quad (3.14)$$

so here $\xi_{ij} = 1$, $\theta_{ij} = \ln(\frac{p_{ij}}{1-p_{ij}})$, $b(\theta_{ij}) = \ln(1 + \exp(\theta_{ij}))$, and $c(X_{ij}, \xi_{ij}) = 0$. Also here $K = 2$, $p = 2$, $n = 29$, and $M = 3$. Using Lemmas 1, 2, and 3, I obtain the conditional likelihood for the whole data:

$$\begin{aligned} L_c &\propto \prod_{i=1}^n \{ \exp\{\theta_{i1}^* X_{i1} - \ln(1 + \exp(\theta_{i1}^*))\} \\ &\times \exp[\sum_{j=2}^{M+1} \{\theta_{ij} X_{ij} - \ln(1 + \exp(\theta_{ij}))\}] \\ &\times \frac{\exp(\mathbf{h}_i^T \boldsymbol{\beta}_{11} Z_{i1}^{(1)} + \mathbf{h}_i^T \boldsymbol{\beta}_{12} Z_{i1}^{(2)}) \times \frac{(1 + \exp(\theta_{i1}^*))}{(1 + \exp(\theta_{i1}))}}{\sum_{j=1}^{M+1} \exp(\mathbf{h}_i^T \boldsymbol{\beta}_{11} Z_{ij}^{(1)} + \mathbf{h}_i^T \boldsymbol{\beta}_{12} Z_{ij}^{(2)}) \times \frac{(1 + \exp(\theta_{ij}^*))}{(1 + \exp(\theta_{ij}))}} \}, \end{aligned} \quad (3.15)$$

where $\theta_{ij}^* = \theta_{ij} + \mathbf{h}_i^T \boldsymbol{\beta}_2$. Since the value of k (i.e., disease type) is completely determined by knowing the stratum one can omit the subscript k for θ_{ijk}^* in the

above expression; $Z_{is}^{(1)}$, and $Z_{is}^{(2)}$ denote the observed value of the two covariates UI, and LWT for the s -th subject in the i -th stratum respectively.

Our analysis is based on normal priors centered at zero with large variances for all the regression parameters. In instances (such as mine) when prior elicitation is not possible, these priors usually lead to posteriors relying more heavily on the data and protect against model failures. In many real applications, the practitioner may have a more precise knowledge about the sign and magnitude of the relative risk parameters, and can suitably change the prior if necessary. We conducted a sensitivity analysis with several choices of prior parameters. For the regression parameters, and the normal base measure of the Dirichlet process, I experimented with normal priors centered at zero, and with variances 2, 4, 5, 6, and 9. I used a Gamma prior on the concentration parameter α of the Dirichlet process and ran my analysis with both shape, and size parameter set at 0.5, 1, 2, 4, 10, 40, 100, 200, and with many other possible priors like $G(0.5, 4)$, $G(1, 10)$, $G(10, 40)$, $G(100, 40)$, $G(200, 10)$. I noted that the ultimate numerical estimates are reasonably stable over a varying range of prior parameters.

The results are reported for independent $N(0, 5)$ prior on each component of β_{11} , β_{12} , and β_2 , $N(0, 6)$ as the mixing normal distribution for the Dirichlet process prior, and a $\text{Gamma}(2, 2)$ prior for the mixing parameter α .

I used componentwise Metropolis Hastings algorithm to generate random numbers from the full conditionals of the parameters. For generating observations from the posterior distribution of γ_{0i} , $i = 1, \dots, n$, I used an algorithm proposed by Neal (2000) for simulating observations from posteriors of Dirichlet mixtures for nonconjugate cases. The details of the computation scheme are given in Chapter 2. I ran the chain typically from 7000-10000 iterations, and the inference was based on the last 3000 MCMC runs.

Table 3-1 contains the posterior means, posterior standard deviations, and 95% HPD credible intervals for the parameters of interest under the proposed Bayesian Semiparametric method (BSP), and the parametric Bayes (PBC and PBV) methods as discussed before. In the parametric Bayes methods I used a $N(0,6)$ prior on the constant γ_0 (in PBC), and i.i.d. $N(0,6)$ prior on varying γ_{0i} 's (in PBV). To illustrate the methods in presence of missingness, I deleted 40% of exposure values completely at random (in the sense of Little and Rubin, 1987) and reran the three analyses. The results are presented in Table 3-2.

The full data analysis indicates that smoking of mother is a significant risk factor for low-birth-weight category (category 1) and is not very significant in the *very* low-birth-weight category (category 2). UI on the other hand shows an opposite association, showing significance in category 2, and almost no significance in category 1. LWT does not seem to be a significant covariate in any of the categories. The BSP, and the PBV methods are in closer agreement whereas the PBC estimates show some numerical differences.

Figure 3-1 shows a plot of the variance of the 29 stratum effects in the last 3000 MCMC samples. The average variance is approximately 2.3 showing that there indeed exists variability in the stratum effects. As a result the BSP and PBV methods which account for this variability are in close agreement, whereas the PBC method assuming constant stratum effect differs numerically from these two methods.

For the analysis with 40% missing observations on smoking one notices that the estimates corresponding to smoking in the BSP method comes closer to their full data counterparts even though the inferences are same in all three methods. As one might expect, with 40% missingness, the parameter estimates for smoking lose precision, and the effect of smoking now appears to be not significant in

both categories 1 and 2. Inferences on the other two covariates remain essentially unchanged when compared to full data inferences.

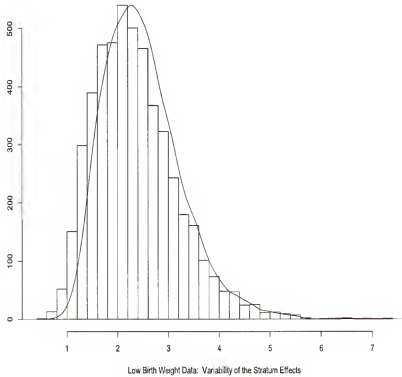


Figure 3-1: Plot of the variability of γ_{0i} 's for the low-birth-weight study example. Estimates of the 29 γ_{0i} 's were collected for each of the last 3000 MCMC samples. Variances of these 29 values were then calculated for each run. The histogram is of these 3000 variance values with a kernel density estimate overlaid on it.

I also analyzed the data after collapsing category 1 and category 2 into only one category (birthweight less than 2500 gms). I carried out the Bayesian analysis for a simple matched case-control data (Sinha et al., 2003), and the usual conditional logistic regression (CLR) analysis (Breslow and Day, 1980). Table 3-3 shows that both BSP and PBV methods assuming varying stratum effects bring out the effect of mother's smoking on having low-birth-weight newborns and

produce very similar results. The PBC and the CLR methods, assuming constant stratum effect are in closer agreement with each other, and they do demonstrate the effect of smoking but not as precisely as the other two methods which allow varying stratum effects. Figure 3-1 again demonstrates the differences in results between these two classes of models. Obviously, without the finer classification into two weight categories, the fact that smoking is not so significant for category 2 and UI is appreciably significant for category 2 cannot be concluded from looking at the overall analysis. Thus the multicategory analysis may render some useful additional information to the practitioner.

3.5 Simulation Study

In the low-birth-weight data one can notice appreciable variability in the stratum effects. In practice, the experimenter may not have a prior idea about the nature of variability among stratum effects and there could be situations where the standard model assumption of constant stratum effect, i.e. $\gamma_{0i} \equiv \gamma_0$ hold. Sinha et al. (2003) contains an example from equine epidemiology where the stratum effects have very small variability. I conducted a simulation study to ascertain the robustness of the BSP method even when variability in the stratum effects is negligible.

In order to simulate a realistic dataset for comparing the BSP, PBC, and PBV methods, I decided to use the low-birth-weight data itself. I generated a hypothetical 1:1 matched data set with 50 strata with one binary exposure variable (corresponding to X , smoking status of mother) and one binary covariate (corresponding to Z , presence of uterine irritability). The true values for β_{11} , β_{12} , β_{21} , and β_{22} are chosen to be 0.20, 1.80, 1.40, and 0.4 respectively, close to the estimates obtained by analyzing the low-birth-weight data by the BSP method as presented in Table 3-1.

In order to elicit values for γ_1 , the coefficient of Z in the natural parameter of the exposure distribution, namely θ_{ij} on Z I ran a logistic regression of $X(\text{SMOKE})$ on $Z(\text{UI})$ using the control sample, and the fitted model turned out to be $\text{logit}(\text{pr}(X_{ij})) = -0.734 + 0.579 Z_{ij}$. Accordingly, in the simulation study I used $\gamma_1 = 0.60$.

I followed the structure of my models as described before for simulating Z , X , and the trinomial variable D , indicating the birth weight category. For the low-birth-weight data occurrence of uterine irritability was reported in approximately 20% of the patients. Thus the completely observed covariate Z was generated first as a Bernoulli variable with success probability 0.20. Second, we generated the trinomial disease variable D . For the i^{th} stratum, one should note that

$$\text{Pr}(D_{i1} = 0 | Z_{i1}, Z_{i2}, D_{i1} + D_{i2} = 1 \text{ or } 2, S_i) = \frac{1}{1 + \frac{Q_{i1}}{Q_{i2}}}$$

where, for $j = 1, 2$,

$$Q_{ij} = \exp(\beta_{11}Z_{ij}) \frac{1 + \exp(\theta_{ij} + \beta_{21})}{1 + \exp(\theta_{ij})} + \exp(\beta_{12}Z_{ij}) \frac{1 + \exp(\theta_{ij} + \beta_{22})}{1 + \exp(\theta_{ij})}. \quad (3.16)$$

I generated a Bernoulli random variable with the above success probability, and if this variable assumed a value 1, the simulated value for D_{i1} was taken to be 0 implying that the first subject in the i -th stratum is a member of the control population. Let, for $k = 1, 2$,

$$p_{ik}^* = \frac{1}{1 + \exp[z_{ik}(\beta_{12} - \beta_{11})] \frac{1 + \exp(\theta_{ik} + \beta_{22})}{1 + \exp(\theta_{ik} + \beta_{21})}}. \quad (3.17)$$

If $D_{i1} = 0$, I determine D_{i2} according to a Bernoulli draw with success probability p_{i2}^* ; set $D_{i2} = 1$ if it results in a success otherwise set $D_{i2} = 2$. If $D_{i1} \neq 0$, set $D_{i2} = 0$, and I determine the value D_{i1} according to a Bernoulli draw with success probability p_{i1}^* ; if this results in a success $D_{i1} = 1$, otherwise $D_{i1} = 2$.

Conditional on the value of D , I proceed to simulate the exposure X . If $D_{ij} = 0$, we generated a binary exposure X_{ij} with success probability given by $\text{logit}(p_{ij}) = \gamma_{0i} + \gamma_1 Z_{ij}$. If $D_{ij} = k$, I generate the binary exposure variable with success probability, $\text{logit}(p_{ij}) = \gamma_{0i} + \gamma_1 Z_{ij} + \beta_{2k}$, $j, k = 1, 2$.

I performed two sets of simulations, one with a constant value of γ_{0i} namely -1.00 the other with a relatively varying set of γ_{0i} 's simulated from a normal distribution with mean -0.5 , and standard deviation 1.5 . I assumed $N(0, 5)$ prior on all the relative risk parameters, and a $\text{Gamma}(2, 2)$ prior for α . In all my simulations I used identical parameters for the normal distribution which is assumed to be prior on γ_0 in PBC, and the i.i.d. prior on γ_{0i} in PBV, and also as the mixing distribution in the Dirichlet Process prior for BSP ($N(0, 6)$ in this case). We replicated the simulation 50 times, generating 50 different datasets, and obtained the parameter estimates by above mentioned methods, and computed their average and MSE. For each replication I also generated data with 30% exposure values missing completely at random and recalculated all the estimates.

The simulation results presented in Tables 3-4 , and 3-5 illustrate that for constant stratum effect, the three methods are comparable with PBV estimates being furthest from the true parameters of interest whereas for varying stratum effect the BSP and PBV methods have a clear edge over the PBC method. Overall BSP seems to be the more robust choice as at the onset of a study one is not sure about the nature of variability in the stratum effects.

3.6 Concluding Remarks

In this Chapter I proposed a semiparametric Bayesian method to analyze matched case-control data with more than one disease state and illustrate the methods with a real example. The simulation results indicate that in presence of stratum variability and missing data the Bayesian semiparametric method

is superior to the parametric Bayesian alternatives. All three methods perform comparably with constant stratum effects.

Our proposed model considers a nondeterministic exposure variable having a probability distribution belonging to the exponential family. Moreover, the distribution of the exposure could be different in different strata. Our model takes into account both discrete and continuous exposure along with possible missingness in the exposure variable. The growing number of stratum effect parameters are modeled in a semiparametric Bayesian way to overcome the inconsistency problems arising out of classical analysis of such matched data. The computations involving a Dirichlet process prior with a mixing normal distribution are done through a suitable MCMC scheme, and enable us to obtain estimates of the parameters of interest. The method could be extended to multiple exposures having an underlying association pattern as well. The general framework is extremely flexible for being used in unorthodox data situations involving missingness and measurement error as well as incorporating widely different types of exposure variables that one may come across in practice.

Table 3-1: Analysis of low-birth-weight data using full dataset with multiple disease states

Logit	Parameter	BSP			PBC			PBV		
		Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD
1	SMOKE	1.42	0.60	(0.33, 2.72)	1.26	0.56	(0.25, 2.50)	1.48	0.65	(0.26, 2.08)
	LWT	-0.86	1.39	(-3.78, 1.81)	-1.03	1.35	(-3.58, 1.86)	-0.73	1.36	(-3.40, 2.01)
	UI	0.15	0.67	(-1.27, 1.46)	0.10	0.67	(-1.19, 1.52)	0.18	0.67	(-1.14, 1.52)
2	SMOKE	Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD region
	LWT	0.37	0.83	(-1.35, 2.05)	0.23	0.66	(-1.10, 1.54)	0.38	0.85	(-1.30, 2.17)
	UI	-0.52	1.61	(-3.76, 2.52)	-0.55	1.59	(-3.73, 2.41)	-0.55	1.62	(-3.65, 2.79)
		1.81	0.83	(0.18, 3.51)	1.78	0.83	(0.30, 3.59)	1.81	0.87	(0.27, 3.72)

BSP stands for Bayesian semiparametric method whereas PBC and PBV stand for parametric Bayes methods assuming constant, and varying stratum effects respectively.

Table 3-2: Analysis of low-birth-weight data after deleting 40% observations on smoking completely at random

Logit	Parameter	BSP			PBC			PBV		
		Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD region
1	SMOKE	0.86	0.88	(-0.88, 2.55)	0.55	0.78	(-0.84, 2.18)	0.56	0.86	(-0.71, 2.49)
	LWT	-0.92	1.31	(-3.63, 1.51)	-1.03	1.34	(-3.50, 1.83)	-1.01	1.33	(-3.57, 1.75)
	UI	0.19	0.69	(-1.11, 1.49)	0.21	0.67	(-1.10, 1.55)	0.20	0.69	(-1.31, 1.48)
2	SMOKE	0.54	1.04	(-1.46, 2.07)	0.13	0.93	(-1.85, 1.88)	0.12	0.93	(-1.10, 1.54)
	LWT	-0.43	1.62	(-3.98, 2.38)	-0.59	1.63	(-3.75, 2.59)	-0.54	1.54	(-3.62, 2.45)
	UI	1.82	0.86	(0.34, 3.65)	1.80	0.83	(0.24, 3.46)	1.87	0.82	(0.43, 3.63)

BSP stands for Bayesian semiparametric method whereas PBC, and PBV stand for parametric Bayes methods assuming constant, and varying stratum effects respectively.

Table 3-3: Analysis of low-birth-weight data after collapsing categories 1 and 2 into a single low-birth-weight category (less than 2500 grams.)

Parameter	BSP			PBC			PBV			CLR	
	Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD region	Estimate	S.E
SMOKE	1.106	0.52	(0.15,2.17)	0.89	0.45	(-0.03,1.78)	1.105	0.54	(0.02,2.25)	0.86	0.45
LWT	-1.07	1.14	(-3.29,1.17)	-1.11	1.14	(-3.28,1.12)	-0.98	1.15	(-3.25, 1.30)	-1.13	1.36
UI	0.85	0.51	(-0.20,1.78)	0.84	0.50	(-0.03,2.03)	0.87	0.49	(-0.13,1.82)	0.85	0.51

BSP stands for Bayesian semiparametric method whereas PBC, and PBV stand for parametric Bayes methods assuming constant and varying stratum effects respectively. CLR stands for usual conditional logistic regression for analyzing matched data.

Table 3-4: Results of the simulation study for multiple disease states, part-I

Full data, fixed $\gamma_{0i} \equiv -1.00$						
Method		β_{11}	β_{12}	β_{21}	β_{22}	γ_1
BSP	Mean	0.23	1.88	1.44	0.40	0.43
	MSE	0.22	6.57	12.50	6.55	36.54
PBC	Mean	0.19	1.92	1.43	0.43	0.51
	MSE	0.57	14.59	14.42	10.41	20.26
PBV	Mean	0.20	1.88	1.43	0.32	0.41
	MSE	0.53	10.56	9.71	15.50	48.36
30% missing data, fixed $\gamma_{0i} \equiv -1.00$						
Method		β_{11}	β_{12}	β_{21}	β_{22}	γ_1
BSP	Mean	0.20	1.90	1.45	0.36	0.44
	MSE	0.17	31.78	50.07	37.83	81.31
PBC	Mean	0.20	1.89	1.47	0.46	0.47
	MSE	4.79	10.05	22.09	11.33	23.98
PBV	Mean	0.20	1.90	1.35	0.28	0.38
	MSE	0.56	9.89	57.05	67.88	83.19

Here "Mean" is the simulated mean, while MSE is the mean squared error $\times 1000$. The true parameter values are $\beta_{11} = 0.20$, $\beta_{12} = 1.80$, $\beta_{21} = 1.40$, $\beta_{22} = 0.40$, and $\gamma_1 = 0.60$. BSP stands for Bayesian semiparametric method whereas PBC, and PBV stand for parametric Bayes methods assuming constant, and varying stratum effects respectively.

Table 3-5: Results of the simulation study for multiple disease states, part-II

Full data, varying $\gamma_{0i} \sim N(-0.5, 2.25)$						
Method		β_{11}	β_{12}	β_{21}	β_{22}	γ_1
BSP	Mean	0.19	1.89	1.43	0.42	0.50
	MSE	0.94	8.20	8.58	10.56	13.47
PBC	Mean	0.20	1.89	1.52	0.48	0.53
	MSE	0.70	8.48	27.98	13.36	97.75
PBV	Mean	0.19	1.92	1.44	0.43	0.51
	MSE	0.65	12.84	11.15	9.07	13.25
30% missing data, varying $\gamma_{0i} \sim N(-0.5, 2.25)$						
Method		β_{11}	β_{12}	β_{21}	β_{22}	γ_1
BSP	Mean	0.20	1.91	1.50	0.44	0.50
	MSE	0.59	11.86	16.96	11.58	19.56
PBC	Mean	0.20	1.92	1.54	0.49	0.53
	MSE	0.46	14.04	30.28	14.68	15.54
PBV	Mean	0.20	1.94	1.49	0.44	0.50
	MSE	0.76	14.31	16.15	14.42	19.35

Here "Mean" is the simulated mean, while MSE is the mean squared error $\times 1000$. The true parameter values are $\beta_{11} = 0.20$, $\beta_{12} = 1.80$, $\beta_{21} = 1.40$, $\beta_{22} = 0.40$, and $\gamma_1 = 0.60$. BSP stands for Bayesian semiparametric method whereas PBC, and PBV stand for parametric Bayes methods assuming constant, and varying stratum effects respectively.

CHAPTER 4
SEMIPARAMETRIC BAYESIAN ANALYSIS OF MATCHED
CASE-CONTROL STUDIES WHEN EXPOSURE VARIABLES
ARE MEASURED WITH ERROR

4.1 Introduction

Errors in measuring exposures can be an important source of bias in epidemiologic studies. In many epidemiologic studies exposure variables can not be observed without error, for instance, measuring a patient's blood pressure, measurement on dietary intake etc. As a result, the estimated odds ratio between a disease and the potential risk factors (relative risk for rare diseases) is subject to bias. This Chapter proposes a method to handle measurement error in exposure variables in matched case-control study when the exposure variables contain partial missing observations. In the likelihood setup, I assume a distribution of the exposure variables in control population given a completely observed covariate and stratum effect. I also assume a prospective model of the disease variable in terms of completely observed covariate, stratum effect, and the exposure variables. A simple additive error structure is assumed for the measurement error. The novelty of my approach is to capture the unmeasured stratum heterogeneity in the distribution of the exposure variables by using a Dirichlet process prior. For all model parameters except the stratum effect parameters, I use parametric prior and they are estimated in a fully Bayesian framework.

The topic of measurement error has drawn considerable attention for decades (Fuller, 1987 and Carroll et al., 1995). For case-control problem, a variety of methods have been proposed to handle measurement error. In most of the situations a fraction of case-control data provide information on response D , and a fallible

surrogate T of the true exposure variable X ; and validation data consist of all the three variables (D, X, T) . Prentice (1982) proposed a method to estimate relative risk parameter in covariate measurement error in failure time data; Carroll and Wand (1991) semiparametrically modeled the distribution of surrogate T and estimated the parameters of logistic regression model. Carroll et al. (1993) proposed a pseudolikelihood method to handle measurement error in case-control studies. They also provided Prentice–Pyke type equivalence of prospective and retrospective analysis of case-control studies in the presence of measurement error. Armstrong et al. (1989) used multivariate discriminant analysis to model the distribution of the exposure variables for matched case-control data. McShane et al. (2001) proposed conditional score procedure for obtaining bias-corrected estimates of log-odds ratio in a matched case-control study. In the Bayesian framework Müller and Roeder (1997) handled measurement error in unmatched case-control studies by assuming a mixture of multivariate normal distribution of the exposure variables. They used a Dirichlet process prior on the mixture components of the distributions. Later on, Seaman and Richardson (2001) extended the Müller–Roeder approach to the situation of any number of categorical covariates and discretized continuous covariates.

Also, as mentioned in previous Chapters, there are plethora of works on missing exposure variable in case-control studies. But to the best of my knowledge, there does not exist any article which simultaneously addresses both missingness and measurement error in an exposure variable in a unified framework.

This Chapter develops a Bayesian approach with a binary disease variable D , completely observed covariate $\mathbf{Z}$ and a surrogate variable $\mathbf{T}$ of true exposure variable $\mathbf{X}$, with partial missing observations. Some additional data, called validity data, has measurement on $\mathbf{T}$ as well as on the true value of the exposure variable $\mathbf{X}$. I start with the usual additive model of error, and assume a multivariate

normal distribution of exposure variable among the controls $D = 0$ given $\mathbf{Z}$, and stratum S . The mean of $\mathbf{X}$ is modeled in terms of completely observed covariates $\mathbf{Z}$, and stratum specific random intercept terms. Further I assume that the errors are nondifferential and follow a multivariate normal distribution. Throughout this Chapter I assume that the surrogate $\mathbf{T}$ is missing at random in the terminology of Little and Rubin (1987). I work with a joint conditional likelihood of D , $\mathbf{T}$, and Δ , the vector of missing value indicator, given the stratum effects, the completely observed covariates $\mathbf{Z}$, and the number of cases in each stratum. I use Dirichlet process prior on the stratum heterogeneity parameters with a normal base measure. As we will see, there is no explicit form of the posterior distribution of the parameters, so I use Gibbs sampler with Metropolis-Hastings algorithm to draw the random numbers from the respective conditional distributions.

The outline of this Chapter is follow. Section 4.2 describe the model, and notation for measurement error while Section 4.3 deals with likelihood, prior and posterior distribution of the parameters for missing value and measurement error in exposure variables. A real data example is presented in Section 4.4 while computational details are considered in Section 4.5 followed by conclusion.

Before concluding this section I highlight some novel features of this Chapter. First, the present Chapter consolidates missingness and measurement error under a unique framework. Second, by using semiparametric Bayesian approach we capture unmeasured stratum heterogeneity in the exposure distribution. Finally, the data example shows the need of undertaking this approach to handle irregular data situation.

4.2 Model and Notation

The study design considers s strata (matched sets), and each stratum consists of 1 case and M controls. Let D be the binary disease indicator. The prospective

disease model for the j^{th} subject in the i^{th} stratum is

$$\text{pr}(D_{ij}|S_i, \mathbf{Z}_{ij}, \mathbf{X}_{ij}) = \frac{\exp\{D_{ij}(\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij})\}}{1 + \exp(\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij})}, \quad (4.1)$$

where $\beta_0(S_i)$ is the stratum specific random intercept, β_1 and β_2 are the log odds ratio parameter corresponding to $q \times 1$ vector of completely observed covariates $\mathbf{Z}$, and $p \times 1$ vector of exposure variables $\mathbf{X}$ respectively.

The usual method of analysis is conditional logistic regression (CLR), and the estimates of the coefficients β_1 , and β_2 are obtained by maximizing the conditional likelihood

$$\begin{aligned} & \text{pr}[D_1, \dots, D_s | \{\mathbf{X}_{ij}; \mathbf{Z}_{ij}, j = 1, \dots, M+1; \sum_{j=1}^{M+1} D_{ij} = 1; S_i\}_{i=1}^s] \\ &= \prod_{i=1}^s \frac{\exp\{\sum_{j=1}^{M+1} D_{ij}(\beta_2^T \mathbf{X}_{ij} + \beta_1^T \mathbf{Z}_{ij})\}}{\sum_{j=1}^{M+1} \exp(\beta_2^T \mathbf{X}_{ij} + \beta_1^T \mathbf{Z}_{ij})} \end{aligned} \quad (4.2)$$

where $\mathbf{D}_i = (D_{i1}, \dots, D_{iM+1})$ is the vector of binary disease variables for the i -th stratum. Without loss of generality one can assume $\mathbf{D}_i = (1, 0, \dots, 0) \forall i$, and then (4.2) reduces to

$$\begin{aligned} & \text{pr}[D_1, \dots, D_s | \{\mathbf{X}_{ij}; \mathbf{Z}_{ij}, j = 1, \dots, M+1; \sum_{j=1}^{M+1} D_{ij} = 1; S_i\}_{i=1}^s] \\ &= \prod_{i=1}^s \frac{\exp(\beta_2^T \mathbf{X}_{i1} + \beta_1^T \mathbf{Z}_{i1})}{\sum_{j=1}^{M+1} \exp(\beta_2^T \mathbf{X}_{ij} + \beta_1^T \mathbf{Z}_{ij})}. \end{aligned} \quad (4.3)$$

When there are measurement errors in the exposure variables $\mathbf{X}$, tests, and estimates of β_2 derived from the above conditional likelihood are biased. Armstrong et al. (1989) and McShane et al. (2001) proposed some methods of how to get bias corrected estimates of the log-odds ratio parameters in the presence of measurement error. But their methods are also inefficient as CLR in the presence of missing observations on the exposure variables. Naive application of these method discard an entire stratum if the case in that stratum is not completely observed. Also, these methods ignore completely observed subjects, if no matching case or

controls are completely observed, in addition to incompletely observed subjects. Moreover, the above mentioned methods of handling measurement error produce biased estimate of the parameters unless the data are missing completely at random (MCAR). In the following Section, I propose a Bayesian method to deal with both missing values and measurement error in exposure variables.

4.3 Likelihood, Priors, and Posteriors

Suppose that the set of exposure variables $\mathbf{X}$ is measured with error, while the set of other covariates $\mathbf{Z}$ is measured without error. Let $\mathbf{T}$ be the error prone observation for $\mathbf{X}$, that is $\mathbf{T}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij}$, $j = 1, \dots, M+1$; $i = 1, \dots, s$; and conditional on $\mathbf{X}$, D is independent of $\mathbf{T}$.

I assume that in the control population the conditional distribution of the exposure variables follow a multivariate normal distribution, i.e.,

$$[\mathbf{X}_{ij}|D_{ij} = 0, \mathbf{Z}_{ij}, S_i] \sim N_p(\boldsymbol{\theta}_{ij}, \mathbf{D}_x), \quad (4.4)$$

where $\boldsymbol{\theta}_{ij} = \Lambda_{0i} + \Lambda_1^T \mathbf{Z}_{ij}$. Here Λ_{0i} captures the unexplained stratum heterogeneity in the exposure distribution which is not measured through $\mathbf{Z}$. Further I assume the nondifferential error model (as described in Carroll and Stefanski, 1994)

$$\mathbf{U}_{ij} \sim N_p(\mathbf{C}, \Sigma_u) \quad \forall i, j. \quad (4.5)$$

The first step is to write the full likelihood handling the measurement error in exposure variables. Using (4.4) in Lemma 1 of Chapter 2 one obtains

$$\begin{aligned} \frac{\text{pr}(D_{ij} = 1|S_i, \mathbf{Z}_{ij})}{\text{pr}(D_{ij} = 0|S_i, \mathbf{Z}_{ij})} &= \int \exp\{\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij}\} \\ &\quad \times \frac{1}{|\mathbf{D}_x|^{\frac{1}{2}}} \exp\{(\mathbf{X}_{ij} - \boldsymbol{\theta}_{ij}) \mathbf{D}_x^{-1} (\mathbf{X}_{ij} - \boldsymbol{\theta}_{ij})\} d\mathbf{X}_{ij} \\ &= \exp[\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \boldsymbol{\theta}_{ij} + \frac{1}{2} \beta_2^T \mathbf{D}_x \beta_2] \end{aligned} \quad (4.6)$$

Using Lemma 2 of Chapter 2 one can easily obtain the distribution of $\mathbf{X}_{ij}$ in the case population as,

$$[\mathbf{X}_{ij}|D_{ij} = 1, \mathbf{Z}_{ij}, S_i] \sim N(\boldsymbol{\theta}_{ij} + \mathbf{D}_x\boldsymbol{\beta}_2, \mathbf{D}_x). \quad (4.7)$$

Now using additive error structure, (4.4), (4.5), and (4.7) one obtains the distribution of $\mathbf{T}$ in control and case population.

$$[\mathbf{T}_{ij}|D_{ij} = 0, \mathbf{Z}_{ij}, S_i] \sim N(\mathbf{C} + \boldsymbol{\theta}_{ij}, \mathbf{D}_x + \mathbf{D}_u) \quad (4.8)$$

$$[\mathbf{T}_{ij}|D_{ij} = 1, \mathbf{Z}_{ij}, S_i] \sim N(\mathbf{C} + \boldsymbol{\theta}_{ij} + \mathbf{D}_x\boldsymbol{\beta}_2, \mathbf{D}_x + \mathbf{D}_u). \quad (4.9)$$

For each subject, missing value in the vector of exposure variables $\mathbf{X}_{ij}$ could occur in $K = 2^p$ many ways. For example, if $\mathbf{X}_{ij}$ has two components (say X_{1ij} and X_{2ij}) possible pattern of missingness are (i) both X_{1ij} and X_{2ij} are missing (ii) X_{1ij} is missing and X_{2ij} is observed, (iii) X_{2ij} is missing and X_{1ij} is observed, and (iv) none of X_{1ij} and X_{2ij} is missing. Let δ^k , $k = 1, \dots, K$ represent the indicator variable corresponding to each pattern of missingness. Therefore, for each subject one observes a vector of indicator variables, $\boldsymbol{\Delta}_{ij} = (\delta_{ij}^1, \delta_{ij}^2, \dots, \delta_{ij}^K)$ which takes value 1 in exactly one position and zero in all other positions. I assume that the missingness patterns are lexicographically ordered, i.e., $\delta_{ij}^1 = 1$ when the missingness pattern is $(0, 0, \dots, 0)$ (denoting all exposures are missing), and $\delta_{ij}^K = 1$ if missingness pattern is $(1, 1, \dots, 1)$, i.e., the exposure vector is completely observed.

Further assume that $p_0^k(\mathbf{X}_{ij})$ and $p_1^k(\mathbf{X}_{ij})$ denote the joint distribution of the observed components of $\mathbf{X}_{ij}$ corresponding to control and case population, for $k = 1, 2, \dots, K$. When all the components of $\mathbf{X}_{ij}$ are missing I set $p_0^1(\mathbf{X}_{ij}) = 1$ and $p_1^1(\mathbf{X}_{ij}) = 1$. Similarly one can define $p_0^k(\mathbf{T}_{ij})$ and $p_1^k(\mathbf{T}_{ij})$ for the surrogate variable $\mathbf{T}$.

Using (4.6)–(4.9) one can compute the full conditional likelihood conditioned on $\sum_{j=1}^{M+1} D_{ij} = 1$. Without loss of generality one can assume that $D_{i1} = 1$, and $D_{ij} = 0, j = 2, \dots, M+1$. Then the likelihood is

$$\begin{aligned}
L_c &= \prod_{i=1}^s \left\{ \text{pr}(D_{i1} = 1, D_{i2} = \dots = D_{iM+1} = 0 | S_i, \{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} D_{ij} = 1) \right. \\
&\quad \times p(\mathbf{T}_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = 1) \times \prod_{j=2}^{M+1} p(\mathbf{T}_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) \left. \right\} \\
&\propto \prod_{i=1}^s \left\{ \frac{\exp[\beta_1^T \mathbf{Z}_{i1} + \beta_2^T \boldsymbol{\mu}_{i1}^*]}{\sum_{r=1}^{M+1} \exp[\beta_1^T \mathbf{Z}_{ir} + \beta_2^T \boldsymbol{\mu}_{ir}^*]} \times \sum_{k=1}^{2^p} \delta_{i1}^k p_1^k(\mathbf{T}_{i1}) \times \prod_{j=2}^{M+1} \sum_{k=1}^{2^p} \delta_{ij}^k p_0^k(\mathbf{T}_{ij}) \right\}
\end{aligned} \tag{4.10}$$

where $\boldsymbol{\mu}_{i1}^* = \Lambda_{0i} + \Lambda_1^T \mathbf{Z}_{i1} + \mathbf{D}_x \boldsymbol{\beta}_2$ and $\boldsymbol{\mu}_{ij} = \Lambda_{0i} + \Lambda_1^T \mathbf{Z}_{ij}$.

The above likelihood involves $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \Lambda_1, \mathbf{D}_x, \mathbf{C}$, and $\mathbf{D}_u$ and a set of nuisance parameters $\Lambda_{0i}, i = 1, 2, \dots, s$ which grows with direct proportion to the sample size. Hence maximum likelihood estimates of the parameters be potentially inconsistent. I plug in the estimates of $\mathbf{C}$ and $\mathbf{D}_u$ obtained from a validity data into the likelihood (4.10). I use diffuse normal prior for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \Lambda_1$, inverse gamma prior on the variance components of $\mathbf{X}$ and uniform $(0, 1)$ prior on the correlation coefficient between any two exposure variables. Finally, I assume that $\Lambda_{0i} \stackrel{iid}{\sim} G$, and $G \sim DP(\alpha G_0)$, Dirichlet process prior with concentration parameter α , and base measure G_0 . Because in our data analysis p equals 2, I took G_0 as a bivariate normal distribution. All the parameters were estimated by Markov chain Monte Carlo technique and computation details are given in Section 4.5.

If there is no missing observation i.e., $\boldsymbol{\Delta}_{ij} = (0, \dots, 0, 1) \forall i, j$, one can work with the following likelihood. The latter does not involve any nuisance parameters. This is in contrast to proposed in Armstrong et al. (1989).

$$L_c = \prod_{i=1}^s \left\{ \text{pr}(D_{i1} = 1, D_{i2} = \dots = D_{iM+1} = 0 | S_i, \{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} D_{ij} = 1) \right\}$$

$$\begin{aligned}
& \times \prod_{j=2}^{M+1} p(\mathbf{T}_{ij} - \mathbf{T}_{i1} | S_i, \mathbf{Z}_{ij}, \mathbf{Z}_{i1}, D_{i1} = 1, D_{ij} = 0) \Bigg\} \\
& \propto \prod_{i=1}^s \left\{ \frac{1}{1 + \sum_{r=2}^{M+1} \exp[\beta_1^T (\mathbf{Z}_{ir} - \mathbf{Z}_{i1}) + \beta_2^T (\boldsymbol{\mu}_{ir} - \boldsymbol{\mu}_{i1}^*)]} \right. \\
& \quad \times \prod_{j=2}^{M+1} \frac{1}{|2(\mathbf{D}_x + \mathbf{D}_u)|^{\frac{1}{2}}} \exp\left[-\frac{1}{4}(\mathbf{T}_{ij} - \mathbf{T}_{i1} - \Lambda_1^T (\mathbf{Z}_{ij} - \mathbf{Z}_{i1}) + \mathbf{D}_x \beta_2)^T \right. \\
& \quad \left. \left. (\mathbf{D}_x + \mathbf{D}_u)^{-1} (\mathbf{T}_{ij} - \mathbf{T}_{i1} - \Lambda_1^T (\mathbf{Z}_{ij} - \mathbf{Z}_{i1}) + \mathbf{D}_x \beta_2) \right) \right\} \quad (4.11)
\end{aligned}$$

4.4 Example: Colon Cancer

I apply the proposed method to a case-control data from a colon cancer study in Canada (see Miller et al., 1983) where 171 male cases of colon cancer were individually matched by age and neighborhood of residence to 171 controls. Height and weight are measured for each study individual apart from their measurement on calories, protein, total fat, dietary fat, and R carotin. The goal of the study is to see the effect of the two continuous exposure variables such as total calories(X_1) and fiber intake(X_2) on the risk of colon cancer. An interesting feature of this dataset is that reliability and validity studies reveal that the measurement on the exposure variables are subject to a correlated measurement error structure. Armstrong *et al.* dealt with the issue of measurement error in this dataset. I propose a Bayesian alternative for correcting for the measurement error and then model the association of the adjusted exposures in terms of completely observed covariates such as height and weight of the subject. I also study the effect of missingness in this situation by introducing artificial missingness in the exposures. This example accounts for interesting unorthodox data scenarios that one may have at hand and addresses the problem of missingness, measurement error and association modeling for a set of continuous exposures under the same framework.

There was an initial study which investigated the validity of the data by comparing dietary histories reported by 16 healthy volunteers with detailed weighed food records kept by the volunteers' spouse. The spousal record were considered as a gold standard. The sample mean and dispersion of the difference between these two records served as a direct estimate of the parameters of the distribution of the measurement error. Therefore the estimate of the mean of $\mathbf{U}_{ij}$ namely, $\mathbf{C}$ and dispersion matrix $\mathbf{D}_u$, are obtained from this validity data. I consider height and weight as two covariates.

Sample means of recorded and reported intakes for total calories and fiber from the validity data were $\bar{\mathbf{T}} = (32.0, 31.9)$ and $\bar{\mathbf{X}} = (25, 20.4)$; their mean difference serves as an estimate of $\mathbf{C}$, namely, $\hat{\mathbf{C}} = (6.9, 11.5)$. The moment estimate of $\mathbf{D}_u$ from the validity data is,

$$\hat{\mathbf{D}}_u = \begin{pmatrix} 56.6 & 35.1 \\ 35.1 & 70.1 \end{pmatrix}.$$

and consequently, $\mathbf{D}_u$ is replaced by $\hat{\mathbf{D}}_u$ in the likelihood in (4.10). I put prior on the components of $\mathbf{D}_x$ in the fully Bayesian analysis. I used $\text{IG}(4.2, 100)$, and $\text{IG}(4, 100)$ priors for $\sigma_{x_1}^2$, and $\sigma_{x_2}^2$ respectively, and uniform prior on the correlation coefficient ($\rho_{x_1 x_2}$) between total calories(X_1), and fiber intake (X_2).

For the proposed analysis I assumes a bivariate normal distribution for the exposure variables in the control population. I used $\text{Normal}(0, 10)$ prior for all the regression parameters and log-odds ratio parameters, and a bivariate normal distribution $\text{BVN}(0, 0, 10, 10, 0.5)$ for the base measure of the Dirichlet process (DP) prior. With the same ideas of previous Chapters, here I consider again two situations: (i) when stratum effect is constant on the distribution of exposure variables, i.e., $\Lambda_{0k} = \Lambda_0$, and use a bivariate normal prior on this parameter, and (ii) when the stratum effect on the distribution of the exposure variable are assumed to vary across strata. In this situation, I use DP prior on G . Since G

is almost surely discrete, the DP prior allows equality of some of the stratum effects. The number of distinct stratum effect depends on the value of α . For the ease of notation, the first, and the second cases are referred to as PBC and BSP respectively. Table 4–1 contains the posterior means, posterior standard deviation and 95% HPD credible intervals for the parameters under the proposed analysis along with the naive analysis of the data without considering measurement error, and the measurement error corrected method in Armstrong et al. (1989) abbreviated as BA.

The naive uncorrected analysis suggests a significant positive association of the disease and total calories, and negative association between the disease and total fiber intake. These association become strong when I consider measurement error in the predictor variables.

I generated 10% missing observation completely at random observation in each of the exposure variable, and reran all the analyses. The results are given in table 4.6. The results are fairly clear. The first point to note that for both situations PBC, and BSP have smaller standard errors for calory than that of BA. For partially missing data the estimates in BA get changes by 50% (Calory), and 76% (Dietary Fiber) which is much higher than that of PBC and BSP. After correction for measurement error, PBC and BSP shows significant negative association between colon cancer, and dietary fiber whereas such effect is absent in BA's analysis. Finally, one can see that in all situations BSP method produces efficient estimate of the parameters in terms of standard errors. In the presence of missing observation, BSP outperforms all other methods and has a clear edge over the other methods as the percentage of missing observation increases in the dataset.

4.5 Computational Details

The basic computational scheme is the same as in the previous Chapters. I highlight primarily the differences that I encounter for the present modeling.

Note that apart from the regression parameter Λ_1 , and the log-odds ratio parameters β_1 and β_2 , I have another set of parameters through D_x . At every cycle, value of each of the parameters are updated by Metropolis-Hastings algorithm and finally with all four updated values of β_1 , β_2 , Λ_1 , and D_x I move to the iteration steps for Λ_{0i} , $i = 1, \dots, s$.

Generating observations from conditional of Λ_{0i} : A key feature of the Dirichlet process is its discreteness, which in my context implies that the Λ_{0i} , $i = 1, \dots, s$ concentrate on a set of some $\kappa \leq s$ distinct values. First of all, the random measure G with Dirichlet process prior has prior mean $E(G) = G_0$, and α the precision parameter determines the closeness of G around G_0 . The larger the α , the closer will G tend to be to G_0 .

Note that the prior of Λ_{0i} conditional on $\Lambda_{0(-i)}$ is,

$$\Lambda_{0i} | \Lambda_{0(-i)} \sim \frac{1}{s + \alpha - 1} \sum_{j \neq i} \delta_{\Lambda_{0j}}(\cdot) + \frac{\alpha}{s + \alpha - 1} G_0(\cdot) \quad (4.12)$$

where $\Lambda_{0(-i)} = (\Lambda_{01}, \dots, \Lambda_{0i-1}, \Lambda_{0i+1}, \dots, \Lambda_{0s})$, the vector of all but i^{th} stratum effect parameter. Note that as $\alpha \rightarrow \infty$, and $\alpha \rightarrow 0$ cluster size become $\kappa = s$, and $\kappa = 1$ respectively. The intuitive idea behind this is that whenever Λ_{0i} comes from G_0 , it will be distinct from the rest, and if $\alpha = 0$ the second term of (4.12) vanishes, and all the Λ_{0i} come from the same cluster. Note that asymptotically, as $s \rightarrow \infty$, the cluster size has a prior mean $\alpha \log(s)$.

When (4.12) combined with the likelihood, this yields the following conditional distribution

$$\Lambda_{0i} | \Lambda_{0(-i)}, y_i \sim \sum_{j \neq i} q_{i,j} \delta(\Lambda_{0j}) + r_i H_i \quad (4.13)$$

Here

$$q_{i,j} = b L_c(y_i, \Lambda_{0j})$$

$$r_i = b\alpha \int L_c(y_i, \theta) G_0(\theta)$$

where y_i is the observed data in the stratum i , and b is such that $\sum_{j \neq i} q_{i,j} + r_i = 1$. Now the Gibbs sampling is feasible if r_i or H_i have a tractable form which is possible if G_0 is conjugate prior. Since G_0 is nonconjugate so I use Metropolis-Hastings algorithm to update each values of Λ_{0i} .

To illustrate the computational steps, let us fix our attention to a particular stratum, say, stratum i . In my data analysis I will take G_0 as a bivariate normal distribution with mean $\zeta_0 = 0$ and variance-covariance matrix

$$\Sigma_0 = \begin{pmatrix} 10 & 5 \\ 5 & 10 \end{pmatrix}.$$

I proceed in the following way:

Step 1. Start with some reasonable values of Λ_{0i} , $i = 1, \dots, s$. The current state of the Markov chain consists of these s values, namely, $(\Lambda_{01}, \dots, \Lambda_{0i}, \dots, \Lambda_{0s})$.

Step 2. Draw a candidate value Λ_{0i}^* from the distribution of $\Lambda_{0i} | \Lambda_{0(-i)}$, namely, from,

$$(\alpha + s - 1)^{-1} \sum_{j \neq i}^s \delta_{\Lambda_{0j}}(\cdot) + \{\alpha / (\alpha + s - 1)\} N_2(\zeta_0, \Sigma_0),$$

There are several ways of generating observations from the above mixture distribution, which essentially reflects that for the i -th stratum effect you either get a new distinct value from the normal component of the mixture with probability $\alpha / (\alpha + s - 1)$, otherwise it is drawn with equal probability from the current set of $(s - 1)$ entries of $\Lambda_{0(-i)}$, the vector corresponding to the other $(i - 1)$ stratum effect parameters.

Step 3. Compute the acceptance probability $a(\Lambda_{0i}^*, \Lambda_{0i}) = \min\{1, L_c(\Lambda_{0i}^* | \cdot) / L_c(\Lambda_{0i} | \cdot)\}$, where $L_c(\Lambda_{0i}^* | \cdot)$, and $L_c(\Lambda_{0i} | \cdot)$ are the conditional likelihoods as given in (4.10) evaluated at Λ_{0i}^* , and Λ_{0i} , respectively, with all other parameters held fixed at their current values.

Step 4. Set the new value of Λ_{0i} to Λ_{0i}^* with the above probability.

Step 5. Repeat Steps 2-4 R times. I chose $R = 5$ in all my computations.

Consider the last of these R updates of Λ_{0i} as the current value of the i -th stratum effect parameter. I repeat the above steps for all s stratum effect parameters.

The full conditional for α is the same as described in the previous Chapter. To be specific, for a $G(a, b)$ prior on α , one has

$$\pi(\alpha|\eta, \kappa) = pG(a + \kappa, b - \log(\eta)) + (1 - p)G(a + \kappa - 1, b - \log(\eta))$$

where $p = (a + \kappa - 1) / [a + \kappa - 1 + s\{b - \log(\eta)\}]$. Here $\eta \sim B(\alpha + 1, s)$ and κ be the number of distinct Λ_{0i} .

4.6 Concluding Remarks

To summarize the finding of this Chapter, one should note that this is the first attempt to address measurement error and missingness in exposure variables in matched case-control studies in a unified framework. Also, here I consider multiple continuous exposures with some association structure and account for unmeasured stratum heterogeneity on exposure distributions, which is not measured through observed covariates.

Table 4-1: Analysis of colon cancer data using full dataset

Parameter	BSP			PBC		
	Mean	Sd	HPD region	Mean	Sd	HPD region
Height	0.33	0.85	(-1.29 ,2.08)	0.46	0.81	(-1.11,2.04)
Weight	0.38	0.89	(-1.39 ,2.14)	0.49	0.79	(-1.14,2.0)
Calory	0.06	0.026	(0.007,0.125)	0.051	0.019	(0.01,0.08)
Dietary Fiber	-0.068	0.026	(-0.12,-0.02)	-0.032	0.014	(-0.05,-0.01)

Parameter	CLR		BA	
	Mean	Sd	Mean	Sd
Height	0.39	0.89	*	*
Weight	0.57	0.86	*	*
Calory	0.042	0.014	0.10	0.034
Dietary Fiber	-0.022	0.012	-0.030	0.019

BSP stands for Bayesian semiparametric method whereas PBC stand for parametric Bayes methods assuming constant stratum effects; and CLR, and BA stand for conditional logistic regression and method proposed in Armstrong et al. (1989) respectively.

Table 4-2: Analysis of colon cancer data with 10% artificial missing observation in each of the exposure variables

Parameter	BSP			PBC		
	Mean	Sd	HPD region	Mean	Sd	HPD region
Height	0.33	0.84	(-1.35,1.85)	0.49	0.82	(-1.15,1.99)
Weight	0.36	0.78	(-1.14,1.84)	0.58	0.79	(-1.08,2.08)
Calory	0.047	0.032	(0.001,0.13)	0.069	0.034	(0.02,0.15)
Dietary Fiber	-0.057	0.028	(-0.12,-0.01)	-0.030	0.014	(-0.06,-0.02)

Parameter	CLR		BA	
	Mean	Sd	Mean	Sd
Height	2.14	1.34	*	*
Weight	0.028	1.99	*	*
Calory	0.068	0.022	0.15	0.056
Dietary Fiber	-0.055	0.018	-0.053	0.022

BSP stands for Bayesian semiparametric method whereas PBC stand for parametric Bayes methods assuming constant stratum effects; and CLR, and BA stand for conditional logistic regression and method proposed in Armstrong et al. (1989) respectively.

CHAPTER 5

MODELLING ASSOCIATION AMONG MULTIVARIATE EXPOSURES IN CASE-CONTROL STUDIES

5.1 Introduction

There are many instances where the exposure variables themselves may be associated among themselves. One such association model is considered in Chapter 4 dealing with multivariate continuous exposure variables. We consider two other association schemes in this Chapter: (i) two exposure variables with a bivariate binary distribution and (ii) two correlated exposure variables, one binary and the other continuous. I propose a full likelihood based approach by a parametric modeling of the multivariate exposure distribution incorporating their association. For matched case-control studies, unexplained stratum heterogeneity may be present. Our model incorporates this stratum heterogeneity on the parameters involved in the exposure distribution. Since the model and the likelihood turn out to be of a complex form with a growing number of parameters, the parameters are estimated in a fully Bayesian framework by using the Markov chain Monte Carlo technique. Also, our model accomodates partial missingness in the exposure variables.

Outside the case-control context there has been extensive amount of work on modeling association among multivariate exposures. Zhao and Prentice (1990) proposed a pseudolikelihood method for the estimation of parameters of a quadratic exponential family in terms of the marginal means and pairwise correlation. Lang et al. (1999) proposed association modeling in multivariate categorical response data. They formulated generalized log-linear models that simultaneously model the association structure as well as the marginal distributions of the responses

and outlined how to obtain the maximum likelihood estimates of the parameters. Ekholm et al. (2000) proposed association models for multivariate binary data through latent variables and Markov type dependence structure. But so far, to the best of my knowledge, none of the models have been extended to the case-control domain.

We reemphasize that one important feature of this Chapter is that it touches on is simultaneous modeling of the joint distribution of categorical and continuous exposures in a case-control context. There has been some relatively recent Bayesian work involving modeling the marginal joint distribution of the exposures. Müller and Roeder (1997) considered a novel semiparametric model for the joint distribution of continuous exposures in presence of measurement error. They assumed a prospective model for the disease status and a joint marginal distribution for the exposures. The evaluation of the retrospective likelihood involves computing the marginal disease probabilities which, in turn, require an integration with respect to the exposure distribution. Computation of this integral becomes hard in a high-dimensional exposure space. Seaman and Richardson (2001) adapted the Müller–Roeder approach for several categorical exposure variables. To avoid the high-dimensional integration, they advocated converting continuous exposures into categorical ones which certainly entails some loss of information. As one should note, in the formulation of the likelihood as presented in this Chapter, I avoid evaluation of the marginal prevalence of disease and consequently, do not need to perform a high dimensional numerical quadrature.

Müller et al. (1999) adopt a fully retrospective model for analyzing case-control data using the joint distribution of a mixed set of quantitative and categorical variables. They assumed a mixture of normal model for the quantitative covariates and conditional on the quantitative covariates, a multivariate probit

model for the categorical covariates. Our formulation of the likelihood and association modeling is different from their proposal. First, I do assume a prospective model and only need to assume the distribution of the exposures in control population in order to accommodate partial missingness in the exposure variables. Müller et al. took a fully retrospective route and thus had to assume the distribution of exposure in both case and control populations. I model the exposure distribution related parameters via a completely observed covariate which is not the case for Müller et al. (1999). Our work focuses on matched case-control studies and accounts for stratum heterogeneity on the exposure distribution whereas Müller et al. considered unmatched studies where there is no need to model such unexplained stratum specific effects.

Two datasets are analyzed in the Chapter, each interesting in its own right. First, I consider the association structure for a bivariate binary exposure in the context of the well-known Los Angeles Endometrial Cancer Study (Breslow and Dey, 1980) which contains natural missingness in one of the exposures. Second, I consider the scenario when there is one binary, and one continuous exposure. A case-control dataset on fibrocystic breast disease is (Pastides et al., 1985) considered to illustrate the methodology. One of the significant exposures is partially missing in this dataset.

In the both examples, I find that by exploiting the association among exposures, especially in the presence of missingness, one obtains better estimates with relatively smaller standard errors where there is a real association among the exposures. A small scale simulation study supports this finding.

The rest of the Chapter is organized as follows. In Section 4.2 I describe my model, notations and formulation of the likelihood in presence of missingness in a matched case-control set-up. Section 4.3 is devoted to two different association

models for multiple exposure variables as described above. Each model is accompanied by the corresponding data example illustrating the proposed methodology. Section 4.4 contains a small scale simulation study to indicate the effect of association modeling when association among exposures truly exists and Section 6 is the concluding section with a discussion of the findings and final remarks.

Before concluding this section I highlight some strikingly new features of this article. First, I consider modeling possible association among exposure variables having potential missing observations in a matched case-control set-up. Second, I consider two different association schemes including simultaneous modeling of continuous and categorical risk factors which may be associated. Third, the presence of unexplained stratum heterogeneity on the exposure distribution is accommodated in my model as I am dealing with highly stratified matched studies. Finally, I address unorthodox data scenarios which include missingness, measurement error and modeling the distribution of a mixed set of discrete and continuous exposures in a matched case-control set-up which is often encountered in real studies. The examples reflect such real scenarios and the need for taking these nonstandard data issues into account while formulating my model and methodology.

5.2 Model and Notation

For subject j in matched set i , $j = 1, \dots, M + 1$; $i = 1, \dots, s$, let D_{ij} be a binary indicator for the disease status, namely, $D_{ij} = 1$ for case and $D_{ij} = 0$ for control. For each subject in the dataset one observes a $p \times 1$ vector of exposure variables $\mathbf{X}_{ij}$ and a $q \times 1$ vector of covariates $\mathbf{Z}_{ij}$ which is completely observed. I allow for the possibility that different components of $\mathbf{X}_{ij}$ may be missing for different subjects. Following the notations of Chapter 4, for each subject one observes a vector of indicator variables $\Delta_{ij} = (\delta_{ij}^1, \delta_{ij}^2, \dots, \delta_{ij}^K)$.

I assume that $p(\mathbf{X}_{ij}|S_i, \mathbf{Z}_{ij}, D_{ij} = 0)$ be a distribution of the exposure variable in the control population with respect to a σ -finite dominating measure μ . Note that by modeling the distribution of the $\mathbf{X}$ in terms of $\mathbf{Z}$ one captures stratum heterogeneity measured through $\mathbf{Z}$ but there may still be some unexplained heterogeneity which would be captured through a random intercept term. I assume that γ_1 be a vector of parameters in the distribution of $\mathbf{X}$ corresponding to $\mathbf{Z}$, and γ_0 be a vector of stratum specific parameters.

Assume that the prospective model for the disease given the exposure variables, covariates and stratum effect is

$$\text{pr}(D_{ij} = 1|S_i, \mathbf{X}_{ij}, \mathbf{Z}_{ij}) = H(\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij}), \quad (5.1)$$

where $H(x) = 1/(1 + e^{-x})$; $\beta_0(S_i)$ is the stratum specific intercept term; and β_1 , and β_2 are the two log-odds ratio parameters corresponding to the completely observed covariates, and the exposure variables.

In case of fully observed data, the conditional likelihood, conditioning on $\sum_{j=1}^{M+1} D_{ij} = 1$, eliminates the nuisance parameters $\beta_0(S_i)$ and involves only the log-odds ratio parameters. In the presence of missingness, the classical method of conditional logistic regression is inefficient. In Chapter 2, I proposed a Bayesian semiparametric method for accounting for missing data via modeling of the exposure distribution. The potential drawback of this method is that it does not take into account the underlying association structure among a set of multivariate exposure variables which may prove to be useful when some exposure variables are missing. The intuitive reason for modeling the association is to borrow information regarding the missing exposure from the corresponding observed exposure variables through the association structure.

The basic structure of the joint likelihood in the presence of missingness is

$$\text{pr}(D_{ij}, \mathbf{X}_{ij}, \Delta_{ij} | \mathbf{Z}_{ij}, S_i) = p(\mathbf{X}_{ij} | D_{ij}, \Delta_{ij}, \mathbf{Z}_{ij}, S_i) \times p(\Delta_{ij} | D_{ij}, \mathbf{Z}_{ij}, S_i) \times \text{pr}(D_{ij} | \mathbf{Z}_{ij}, S_i). \quad (5.2)$$

Note that instead of using a retrospective likelihood I work with a joint likelihood of disease variable, exposure variable, and missing value indicator given the completely observed covariates, and the stratum effect. The consistency and asymptotic property of the maximum likelihood estimator obtained from the above likelihood follow from Rathouz (2003).

Following the notations of Chapter 4, and assuming without loss of generality that the first subject in each stratum is a case and rest are controls, by Lemmas 1 and 2 of Chapter 2, one begins with the likelihood

$$\begin{aligned} L_c(\beta_1, \beta_2, \gamma_1, \gamma_0) &= \prod_{i=1}^s \text{pr}(D_{i1} = 1, D_{i2} = 0, \dots, D_{iM+1} = 0, \{\mathbf{X}_{ij}, \Delta_{ij}\}_{j=1}^{M+1} | S_i, \{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \\ &\quad \sum_{r=1}^{M+1} D_{ir} = 1) \\ &\propto \prod_{i=1}^s \left\{ \text{pr}(D_{i1} = 1, D_{i2} = 0, \dots, D_{iM+1} = 0, | S_i, \{\mathbf{Z}_{ij}\}_{j=1}^{M+1}, \sum_{r=1}^{M+1} D_{ir} = 1) \right. \\ &\quad \times p(\mathbf{X}_{i1} | S_i, \mathbf{Z}_{i1}, D_{i1} = 1, \Delta_{i1}) \prod_{j=2}^{M+1} p(\mathbf{X}_{ij} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0, \Delta_{ij}) \left. \right\} \\ &\propto \prod_{i=1}^s \frac{\text{pr}(D_{i1} = 1 | \mathbf{Z}_{i1}, S_i) / \text{pr}(D_{i1} = 0 | \mathbf{Z}_{i1}, S_i)}{\sum_{j=1}^{M+1} \text{pr}(D_{ij} = 1 | \mathbf{Z}_{ij}, S_i) / \text{pr}(D_{ij} = 0 | \mathbf{Z}_{ij}, S_i)} \\ &\quad \times \prod_{i=1}^s \left\{ \sum_{k=1}^{2^p} \delta_{i1}^k p_1^k(\mathbf{X}_{i1}) \times \prod_{j=2}^{M+1} \sum_{k=1}^{2^p} \delta_{ij}^k p_0^k(\mathbf{X}_{ij}) \right\}. \end{aligned} \quad (5.3)$$

The above expression is a generalization of (4.10). The parameters involved in the above likelihood are estimated in a fully Bayesian framework. The details will be provided for each individual case. As one will find, there will be no closed form available for the posterior and any exact Bayesian inference is analytically

intractable. Hence I estimate the model parameters through Markov chain Monte Carlo numerical integration, and in particular, via Gibbs sampling techniques of generating sample from conditional distributions of the different parameters given the remaining parameters and the data. The general computational scheme is similar as given in previous Chapters with the added simplification that we have not used any Dirichlet process prior on the stratum effect parameters and instead have used only independent normal priors. Use of Dirichlet process prior in this context is a part of my future research.

In the following subsections I present the specific forms of the likelihoods and other details for the two association scenarios I consider.

5.2.1 Bivariate Binary Exposure Variable

I first consider the situation with bivariate binary exposure variables $\mathbf{X}_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})$ with corresponding log odds ratio parameter $\beta_2 = (\beta_{21}, \beta_{22})$. Assume the retrospective distribution of the exposure variable in control population to be,

$$p(X_{ij}^{(1)}, X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) = \frac{1}{C_{ij}^{(0)}} \exp(\theta_{ij}^{(1)} X_{ij}^{(1)} + \theta_{ij}^{(2)} X_{ij}^{(2)} + \lambda_{ij} X_{ij}^{(1)} X_{ij}^{(2)}) \quad (5.4)$$

where $\theta_{ij}^{(1)}$, $\theta_{ij}^{(2)}$, and λ_{ij} are modeled as $\theta_{ij}^{(1)} = \gamma_{01i} + \gamma_{11} \mathbf{Z}_{ij}$, $\theta_{ij}^{(2)} = \gamma_{02i} + \gamma_{12} \mathbf{Z}_{ij}$, and $\lambda_{ij} = \gamma_{03i} + \gamma_{13} \mathbf{Z}_{ij}$. $C_{ij}^{(0)}$ is the normalizing constant.

Remark 1.: Note that λ_{ij} is the odds ratio between the two random variables, i.e.,

$$\lambda_{ij} = \frac{p(X_{ij}^{(1)} = 1, X_{ij}^{(2)} = 1 | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) p(X_{ij}^{(1)} = 0, X_{ij}^{(2)} = 0 | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)}{p(X_{ij}^{(1)} = 0, X_{ij}^{(2)} = 1 | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) p(X_{ij}^{(1)} = 1, X_{ij}^{(2)} = 0 | S_i, \mathbf{Z}_{ij}, D_{ij} = 0)} \quad (5.5)$$

and $\theta_{ij}^{(1)}$ could be interpreted as the logit of probability of success of $X_{ij}^{(1)}$ conditioned on $X_{ij}^{(2)} = 0$, i.e., $p(X_{ij}^{(1)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(2)} = 0) = \exp(X_{ij}^{(1)} \theta_{ij}^{(1)}) / \{1 + \exp(\theta_{ij}^{(1)})\}$. Similarly $\theta_{ij}^{(2)}$ could be interpreted as the logit of the probability of success of $X_{ij}^{(2)}$ given that $X_{ij}^{(1)} = 0$.

Using (5.4) in Lemma 1 of Chapter 2 one obtains,

$$\frac{\text{pr}(D_{ij} = 1|S_i, \mathbf{Z}_{ij})}{\text{pr}(D_{ij} = 0|S_i, \mathbf{Z}_{ij})} = \exp\{\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \log(C_{ij}^{(1)}/C_{ij}^{(0)})\}, \quad (5.6)$$

where $C_{ij}^{(0)} = 1 + \exp(\theta_{ij}^{(1)}) + \exp(\theta_{ij}^{(2)}) + \exp(\theta_{ij}^{(1)} + \theta_{ij}^{(2)} + \lambda_{ij})$ and $C_{ij}^{(1)} = 1 + \exp(\theta_{ij}^{(1)} + \beta_{21}) + \exp(\theta_{ij}^{(2)} + \beta_{22}) + \exp(\theta_{ij}^{(1)} + \theta_{ij}^{(2)} + \lambda_{ij} + \beta_{21} + \beta_{22})$.

Using Lemma 2 of Chapter 2 and (5.6) one obtains the exposure distribution in the case population as:

$$p(X_{ij}^{(1)}, X_{ij}^{(2)}|S_i, \mathbf{Z}_{ij}, D_{ij} = 1) = \frac{1}{C_{ij}^{(1)}} \exp(\overline{\theta_{ij}^{(1)} + \beta_{21}X_{ij}^{(1)} + \theta_{ij}^{(2)} + \beta_{22}X_{ij}^{(2)} + \lambda_{ij}X_{ij}^{(1)}X_{ij}^{(2)}}). \quad (5.7)$$

Marginally each component of $\mathbf{X}_{ij}$ is a binary random variable and one can easily obtain the marginal distributions in case and control population. In this situation $\Delta_{ij} = (\delta_{ij}^1, \delta_{ij}^2, \delta_{ij}^3, \delta_{ij}^4)$. Using (5.4), (5.6) and (5.7) we can write the likelihood as

$$L_c \propto \prod_{i=1}^s \left\{ \frac{\exp\{\beta_1^T \mathbf{Z}_{i1} + \log(C_{i1}^{(1)}/C_{i1}^{(0)})\}}{\sum_{j=1}^{M+1} \exp\{\beta_1^T \mathbf{Z}_{ij} + \log(C_{ij}^{(1)}/C_{ij}^{(0)})\}} \times \sum_{k=1}^4 \delta_{i1}^k p_1^k(\mathbf{X}_{i1}) \right. \\ \left. \times \prod_{j=2}^{M+1} \sum_{k=1}^4 \delta_{ij}^k p_0^k(\mathbf{X}_{ij}) \right\}. \quad (5.8)$$

To estimate the parameters I use independent normal prior on each of the parameters and estimate them through Markov chain Monte Carlo integration scheme.

Example: Endometrial Cancer Data

Let us revisit the endometrial cancer data as considered in Chapter 2. Most of the major risk factors for endometrial cancer can be characterized as states of excess unopposed estrogenic stimulation of the endometrium. Several studies have indicated that exogeneous estrogen treatment increases the risk of endometrial cancer (see Schottenfeld and Fraumeni, 1996). Among other factors, obesity may increase the risk of endometrial cancer by altering estrogen metabolism. Hypertension has also been reported to be common among women with endometrial cancer. But several studies controlling for certain confounding factors such as,

parity, recent weight gain or estrogen use, showed no evidence of increased risk of endometrial cancer in presence of hypertension, concluding that the apparent effect of hypertension is in fact manifested through increased levels of obesity and estrogen use. The presence of gall bladder disease may increase the risk of endometrial cancer and that could potentially be due to an association of gall bladder disease with obesity or estrogen replacement therapy. Obesity is a potential exposure with 16% values missing in the dataset. Though there are several other potential risk factors such as, smoking, total dietary fat and alcohol consumption recorded in this dataset, it has been scientifically established that the use of estrogen, depending on the dose level, plays a major role in endometrial cancer disease. With this underlying medical theory in mind, I reanalyzed the LA cancer dataset with obesity and gall bladder disease as two associated binary exposures. Estrogen plays the role of a completely observed covariate Z through which parameters related to the bivariate binary exposure distribution are modeled.

Bivariate binary distribution is assumed for the two exposure variables in the control population. The related parameters $\theta_{ij}^{(1)}$, $\theta_{ij}^{(2)}$ and λ_{ij} 's are modeled in terms of Z (estrogen use) as: $\theta_{ij}^{(1)} = \gamma_{01i} + \gamma_{11}Z_{ij}$, $\theta_{ij}^{(2)} = \gamma_{02i} + \gamma_{12}Z_{ij}$, and $\lambda_{ij} = \gamma_{03i} + \gamma_{13}Z_{ij}$. I used Normal(0,10) priors for each of the parameters. For each of the examples I perform three analyses. The first is my proposed method of Bayesian analysis accounting for the association among the exposure variables (denoted by AM). The second is a Bayesian method following Sinha *et al.* (2003) where exposure variables are assumed to be independent (denoted by IM) and the third one is the usual conditional logistic regression (CLR). For the independence modeling I set the log-odds ratio coefficient, λ_{ij} equal to zero in (5.4).

I carried out the three analyses on the observed data. I reran all the three methods when 30% observations on the exposure variable (the presence of gall

bladder disease) were randomly deleted. I then compared the performances of the three methods in presence and absence of missingness.

To summarize the results, for the observed data, one can notice that use of estrogen and gall bladder disease are both significant risk factors for endometrial cancer. There is very little numerical difference between the estimates and their standard errors obtained by using the AM model and the IM model. In contrast, for CLR one observes a significantly higher value of the standard error. With artificial missingness in gall bladder disease one can notice a significant difference in the estimates, and in the standard errors between AM and IM. AM produces estimates with smaller standard error than IM and the AM estimates are relatively closer to their original observed data counterparts. The explanation lies in the fact that the association between gall-bladder disease, and obesity is exploited in the AM model, yielding more efficient estimates. As expected CLR produces worse estimates in presence of missing exposure variable.

5.2.2 One Binary Exposure and Another Belonging to an Exponential Family

In this section, I consider the modeling of a mixed set of continuous and categorical exposures. Though I write my model with the categorical exposure to be binary, the methods may easily be extended to any categorical exposure.

Suppose $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are two exposure variables for the j^{th} subject in the i^{th} stratum. I assume that $X_{ij}^{(1)}$ is binary and the conditional distribution of $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ belongs to a general exponential family model for all i and j . This implies that the marginal distribution of $X^{(2)}$ belongs to a two component mixture of general exponential family. Thus,

$$p(X_{ij}^{(1)} = 1 | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) = \pi_{ij} \quad (5.9)$$

where $\text{logit}(\pi_{ij}) = \gamma_{0i} + \gamma_1 \mathbf{Z}_{ij}$ and the conditional distributions of $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ are assumed to be

$$p(X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 0) = \exp\left\{\frac{\theta_{0ij}X_{ij}^{(2)} - b(\theta_{0ij})}{\phi_0} + c(X_{ij}^{(2)}, \phi_0)\right\}, \quad (5.10)$$

$$p(X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 1) = \exp\left\{\frac{\theta_{1ij}X_{ij}^{(2)} - b(\theta_{1ij})}{\phi_1} + c(X_{ij}^{(2)}, \phi_1)\right\}, \quad (5.11)$$

where $\theta_{0ij} = \gamma_{02i} + \gamma_2^T \mathbf{Z}_{ij}$ and $\theta_{1ij} = \gamma_{03i} + \gamma_3^T \mathbf{Z}_{ij}$. Note the marginal distribution of $X_{ij}^{(2)}$ is a two component mixture of exponential distribution in both the case and control population.

Remark 2. Note that if $X_{ij}^{(2)}$ is a binary exposure variable with the following parameterization,

$$p(X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 0) = \exp\{\theta_{0ij}X_{ij}^{(2)} - \log(1 + \exp(\theta_{0ij}))\}$$

$$p(X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 1) = \exp\{\theta_{1ij}X_{ij}^{(2)} - \log(1 + \exp(\theta_{1ij}))\}$$

and $p(X_{ij}^{(1)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) = \exp\{\xi_{ij}X_{ij}^{(1)} - \log(1 + e^{\xi_{ij}})\}$, then the joint distribution of $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ is the bivariate binary distribution (5.4) of Section 2.1 with $\theta_{ij}^{(1)} = \xi_{ij} + \log\{(1 + \exp(\theta_{0ij}))(1 + \exp(\theta_{1ij}))\}$, $\theta_{ij}^{(2)} = \theta_{0ij}$, and $\lambda_{ij} = \theta_{1ij} - \theta_{0ij}$.

Using prospective logistic model for the disease status, and (5.9)-(5.11) in Lemma 1 of Chapter 2 one obtains,

$$\begin{aligned} \frac{\text{pr}(D_{ij} = 1 | S_i, \mathbf{Z}_{ij})}{\text{pr}(D_{ij} = 0 | S_i, \mathbf{Z}_{ij})} &= \int \int \exp(\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij} + \beta_2^T \mathbf{X}_{ij}) p(X_{ij}^{(1)}, X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, D_{ij} = 0) \\ &\quad d\mu(X_{ij}^{(1)}) dX_{ij}^{(2)} \\ &= \exp(\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij}) \left\{ (1 - \pi_{ij}) \int \exp(\beta_{22}X_{ij}^{(2)}) p(X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, \right. \\ &\quad \left. X_{ij}^{(1)} = 0, D_{ij} = 0) dX_{ij}^{(2)} \right. \\ &\quad \left. + \pi_{ij} \int \exp(\beta_{21} + \beta_{22}X_{ij}^{(2)}) p(X_{ij}^{(2)} | S_i, \mathbf{Z}_{ij}, X_{ij}^{(1)} = 1, D_{ij} = 0) dX_{ij}^{(2)} \right\} \end{aligned}$$

$$\begin{aligned}
&= \exp(\beta_0(S_i) + \beta_1 \mathbf{Z}_{ij}) \left\{ (1 - \pi_{ij}) \exp\left\{ \frac{b(\theta_{0ij} + \phi_0 \beta_{22}) - b(\theta_{0ij})}{\phi_0} \right\} \right. \\
&\quad \left. + \pi_{ij} \exp\left\{ \frac{b(\theta_{1ij} + \phi_1 \beta_{22}) - b(\theta_{1ij})}{\phi_1} + \beta_{21} \right\} \right\} \\
&= \exp(\beta_0(S_i) + \beta_1^T \mathbf{Z}_{ij}) g_{ij},
\end{aligned} \tag{5.12}$$

where

$$g_{ij} = (1 - \pi_{ij}) \exp\left\{ \frac{b(\theta_{0ij} + \phi_0 \beta_{22}) - b(\theta_{0ij})}{\phi_0} \right\} + \pi_{ij} \exp\left\{ \frac{b(\theta_{1ij} + \phi_1 \beta_{22}) - b(\theta_{1ij})}{\phi_1} + \beta_{21} \right\}, \tag{5.13}$$

and $\mu(\cdot)$ is a counting measure. Using Lemma 2, the marginal distribution of $X_{ij}^{(2)}$ in the case population is again seen to be a mixture of exponential family of distributions. The joint conditional likelihood is

$$L_c \propto \prod_{i=1}^s \left\{ \frac{\exp(\beta_1^T \mathbf{Z}_{i1}) g_{i1}}{\sum_{j=1}^{M+1} \exp(\beta_1^T \mathbf{Z}_{ij}) g_{ij}} \times \sum_{k=1}^4 \delta_{i1}^k p_1^k(\mathbf{X}_{i1}) \times \prod_{j=2}^{M+1} \sum_{k=1}^4 \delta_{ij}^k p_0^k(\mathbf{X}_{ij}) \right\} \tag{5.14}$$

Note that the extension of this method to the situation when conditional distribution of $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ belongs to a multi-parameter generalized exponential family, is straightforward and do not need any extra calculation.

Example: Fibrocystic Breast Disease

This data is part of a large case-control study done in 1979 at two hospitals in New Haven, Connecticut (see Pastides et al., 1985) to assess the risk factors for benign fibrocystic breast diseases. The original data consists of 255 women (cases) aged 20-74 years with biopsy confirmed fibrocystic breast lesions and 790 controls at two hospitals in New Haven, Connecticut during 1979 (Pastides et al.). The purpose of the study was to ascertain whether known risk factors for malignant breast cancer are also established risk factors for this type of benign breast diseases. There are many variables recorded in this dataset including demographic characteristics of the patient, medical history information, history of breast cancer in the family. The fraction of the data used in my analysis consists

of 50 strata each of which contains 1 case and 3 controls. Controls were matched with a case on the basis of their age at the time of interview. The dataset was discussed in Hosmer and Lemeshow (2000, Section 7.8). There are some variables which are naturally missing in the dataset. Cases were found to have a higher level of education, a recent history of medical check-ups and a higher age at first pregnancy. The latter two exposures have natural missingness in the dataset. The established risk factor for this disease is age at menarche (first menstrual period), late age at menopause, late age at first full term pregnancy, postmenopausal obesity, low physical activity and higher dose exposure to ionizing radiation early in life. The data contain several other covariates such as the educational degree, marital status, regular medical checkups, number of stillbirth, number of miscarriages, number of live birth. For the purpose of my analysis I consider education as the completely observed covariate Z . History of medical checkups and age at first pregnancy are considered as two exposure variables, say $X^{(1)}$ and $X^{(2)}$. Z and $X^{(1)}$ are defined specifically as follows:

$$Z_{ij} = \begin{cases} 1 & \text{if the } ij^{th} \text{ subject has a junior college degree or above} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij}^{(1)} = \begin{cases} 1 & \text{if the } ij^{th} \text{ subject had regular medical checkups} \\ 0 & \text{otherwise} \end{cases}$$

I model $\log\{\pi_{ij}/(1 - \pi_{ij})\} = \gamma_{01i} + \gamma_1 Z_{ij}$, and the canonical parameters involved in the distributions of $X^{(2)}$ are modeled as, $\theta_{0ij} = \gamma_{02i} + \gamma_2 Z_{ij}$ and $\theta_{ij} = \gamma_{03i} + \gamma_3 Z_{ij}$. $X^{(2)}$ was re-scaled by a factor of 10. I further assumed that

$$[X_{ij}^{(2)} | S_i, Z_{ij}, D_{ij} = 0, X_{ij}^{(1)} = 0] \sim \text{Normal}(\theta_{0ij}, \sigma_0^2),$$

and

$$[X_{ij}^{(2)} | S_i, Z_{ij}, D_{ij} = 1, X_{ij}^{(1)} = 1] \sim \text{Normal}(\theta_{1ij}, \sigma_1^2).$$

It follows that the marginal distribution of $X^{(2)}$ in both the case and the control population is a mixture of two normal distributions. For each of the parameters β_1

(since Z is a single covariate), β_2 , γ_1 , γ_2 , γ_3 and γ_{01i} , γ_{02i} , γ_{03i} , for $i = 1, 2, \dots, s$ I use independent and fairly diffuse normal priors. The posterior means of the parameters are taken as our estimates. The results are presented in Table 5–3. For the independence model we assume that the marginal distribution of $X^{(2)}$ is a single normal distribution (in general a single distribution belonging in the exponential family) instead of a mixture and carry out the three analysis. Here the important exposure age at first pregnancy had 11% missing values in the original dataset. One can notice some interesting numerical differences between AM and IM estimates with AM estimates reflecting a more pronounced effect of the established risk factors. CLR is less efficient than any of the Bayesian methods.

5.3 Simulation Study

To study the effectiveness of association modeling in the context of matched case-control study I performed a small simulation study. The performance of the association modeling was compared with independence modeling and conditional logistic regression method in two different situations, when two exposure variables are associated, and when they are independent.

To simulate realistic data set I used endometrial cancer data as a prototype. I generated hypothetical 1:1 matched dataset with 100 strata with a covariate Z and two exposure variables $X^{(1)}$ and $X^{(2)}$. The true value of log-odds ratio parameters, such as β_1 , β_{21} , and β_{22} , and other parameters involved in the joint distribution of the exposure variables such as γ_{11} , γ_{12} and γ_{13} were taken as 2.18, 1.11, 0.67, -1.40, 0.19, and 0.44 respectively. These are the estimates of the parameters obtained by analyzing the dataset through association modeling approach. The stratum specific intercept terms γ_{01i} , γ_{02i} , and γ_{03i} were generated from $\text{Normal}(-1.40, 2.37)$, $\text{Normal}(0.42, 2.85)$, and $\text{Normal}(-0.67, 2.95)$ respectively as were obtained from the data analysis. In the original dataset, gall bladder disease was reported only for

19% of the subjects so the covariate Z was generated as Bernoulli random variable with success probability 0.19.

Second, I generated the binary disease indicator D . For the i^{th} stratum, it may be easily noted that,

$$\text{pr}(D_{i1} = 1 | D_{i1} + D_{i2} = 1, Z_{i1}, Z_{i2}, S_i) = \frac{1}{1 + \exp\{\beta_1(Z_{i2} - Z_{i1})\} \frac{C_{i2}^{(1)} C_{i1}^{(0)}}{C_{i2}^{(0)} C_{i1}^{(1)}}} \quad (5.15)$$

where for $j = 1, 2$

$$C_{ij}^{(0)} = 1 + \exp(\theta_{ij}^{(1)}) + \exp(\theta_{ij}^{(2)}) + \exp(\theta_{ij}^{(1)} + \theta_{ij}^{(2)} + \lambda_{ij}),$$

$$C_{ij}^{(1)} = 1 + \exp(\theta_{ij}^{(1)} + \beta_{21}) + \exp(\theta_{ij}^{(2)} + \beta_{22}) + \exp(\theta_{ij}^{(1)} + \theta_{ij}^{(2)} + \lambda_{ij} + \beta_{21} + \beta_{22});$$

and $\theta_{ij}^{(1)} = \gamma_{01i} + \gamma_{11}Z_{ij}$, $\theta_{ij}^{(2)} = \gamma_{02i} + \gamma_{12}Z_{ij}$, and $\lambda_{ij} = \gamma_{03i} + \gamma_{13}Z_{ij}$. I generated a binary random variable with the above specified probability structure and conditional on the fact that the simulated value for D_{i1} is a 1 (i.e., the subject is a case) I generated a binary exposure variable ($X_{i1}^{(1)}$) from a Bernoulli distribution with success probability p_1 where

$$p_1 = \frac{1}{C_{i1}^{(1)}} \exp(\theta_{i1}^{(1)} + \beta_{21}) \{1 + \exp(\theta_{i1}^{(2)} + \lambda_{i1} + \beta_{22})\}.$$

Now the second binary exposure variable ($X_{i1}^{(2)}$) is generated from a Bernoulli distribution whose success probability $p_{2|1}$ depends on the value of $X_{i1}^{(1)}$, and is given by

$$p_{2|1} = \frac{\exp\{\lambda_{i1}X_{i1}^{(1)} + \theta_{i1}^{(2)} + \beta_{22}\}}{1 + \exp\{\lambda_{i1}X_{i1}^{(1)} + \theta_{i1}^{(2)} + \beta_{22}\}}.$$

If the simulated value for D_{i1} is 0 (i.e., the subject is a control) I generate $X_{i1}^{(1)}$ from a Bernoulli distribution with success probability $p_1 = \exp(\theta_{i1}^{(1)}) \{1 + \exp(\theta_{i1}^{(2)} + \lambda_{i1})\} / C_{i1}^{(0)}$; and subsequently $X_{i1}^{(2)}$ is generated from a Bernoulli distribution, conditioned on $X_{i1}^{(1)}$, with the success probability $p_{2|1} = \exp\{\lambda_{i1}X_{i1}^{(1)} + \theta_{i1}^{(2)}\} / (1 + \exp\{\lambda_{i1}X_{i1}^{(1)} + \theta_{i1}^{(2)}\})$. Once I have simulated D_{i1} in the above manner, for the other observation in the i^{th} stratum, $D_{i2} = 1 - D_{i1}$. The corresponding exposure variables are then generated in the above fashion.

I performed two sets of simulations: one when exposure variables are associated, and the other when they are independent (i.e., the log-odds ratio parameter between the two exposure variable $\lambda_{ij} = 0.00 \ \forall i, j$). For the Bayesian methods, (i.e., AM and IM), I used $\text{Normal}(0, 10)$ prior for each of the parameters. I replicated the simulation 50 times, generating 50 different data sets and obtained the parameter estimates by the above three methods. For each replication I created missing data by randomly deleting 20% observations from each of the exposure variables and recalculated all the estimates by the three methods (i.e, AM, IM, and CLR).

The result of the simulation study is fairly clear. When the exposure variables are truly associated and partially missing, AM performs much better than IM and CLR. When the exposure variables are not associated, AM and IM perform comparably and both outperform CLR. There is a clear edge of AM over the other two methods. The relative advantage of using the AM increases as the percentage of missing observation increases in a dataset. AM allows the possibility to borrow information on missing values of an exposure variable from the other observed components of the exposure variable via the association structure and also through the completely observed covariate. But, due to independence assumption, IM is not able to extract information on the missing components of an exposure variable from the observed components of the exposure variable, it can only use the completely observed covariate information. Thus, in the presence of association and missingness IM is less efficient.

5.4 Concluding Remarks

To summarize the finding of this Chapter one can note that this is the first attempt to account for multivariate association among exposures in missing data situations in a matched case-control study. Through my examples I find that the association model outperforms the independence model when one has two

strongly associated exposures and an informative completely observed covariate. I also present a brief discussion of treatment of missingness and measurement error under the same framework. As mentioned in the introduction this model presents a way to deal with categorical and continuous exposures simultaneously. The work presented here for binary exposures can be extended to a general categorical exposure in a straightforward way. The model also accounts for stratum heterogeneity on exposure distributions through a stratum-specific intercept term while modeling the parameters in the exposure distribution. Thus this Chapter offers an ensemble of statistical methods for unorthodox data situations in a matched case-control study.

Table 5-1: Results of the endometrial cancer data for the association modelling

Endometrial Cancer Data				
Method		Estrogen-Use	Gall-Bladder	Obesity
AM	Mean	2.18	1.11	0.67
	s.e.	0.42	0.40	0.39
	Lower HPD	1.48	0.35	-0.24
	Upper HPD	3.15	1.94	1.40
IM	Mean	2.23	1.11	0.65
	s.e.	0.45	0.41	0.41
	Lower HPD	1.42	0.35	-0.20
	Upper HPD	3.20	1.95	1.43
CLR	Mean	1.95	1.26	0.42
	s.e.	0.50	0.44	0.40
	Lower CL	0.98	0.41	-0.36
	Upper CL	2.93	2.11	1.19
With 30% Artificial Missingness on Gall-Bladder				
Method		Estrogen-Use	Gall-Bladder	Obesity
AM	Mean	2.37	1.31	0.75
	s.e.	0.45	0.50	0.40
	Lower HPD	1.56	0.25	-0.28
	Upper HPD	3.32	2.42	1.50
IM	Mean	2.52	1.70	0.64
	s.e.	0.49	0.56	0.42
	Lower HPD	0.66	0.68	-0.22
	Upper HPD	3.45	2.86	1.40
CLR	Mean	1.69	0.87	0.07
	s.e.	0.64	0.58	0.55
	Lower CL	0.45	-0.27	-1.00
	Upper CL	2.94	2.01	1.15

Here “Mean” is the posterior mean, “s.e.” is the posterior standard deviation, “Lower HPD” and “Upper HPD” are the lower and upper end of the HPD region respectively, “Lower CL” and “Upper CL” are the lower and upper end of the confidence limit respectively. There is natural missingness in Obesity

Table 5-2: Secondary model parameter estimates for endometrial cancer data

Parameter		Association Modeling		Independence Modeling	
		Full Data	Missing Data	Full Data	Missing Data
γ_{11}	Mean	-1.40	-1.81	-1.17	-2.07
	S. E.	0.45	0.53	0.35	0.49
γ_{12}	Mean	0.19	0.26	0.22	0.25
	S.E.	0.30	0.31	0.30	0.32
γ_{13}	Mean	0.44	0.33		
	S.E.	0.53	0.63		

Secondary model parameter estimates for association and independence modeling of the Endometrial Cancer Data example. Here “Mean” is the posterior mean, “S.E.” is the posterior standard deviation. Missing data implies the data with 30% artificial missingness on the gall-bladder disease.

Table 5-3: Results of the benign breast disease data

Method		Benign Breast Disease		
		Education	Regular Medical Checkups	Age at First Pregnancy
AM	Mean	-0.62	1.47	1.86
	s.e.	0.43	0.43	0.50
	Lower HPD	-1.40	0.65	0.78
	Upper HPD	0.38	2.42	2.85
IM	Mean	-0.38	1.61	1.50
	s.e.	0.42	0.43	0.52
	Lower HPD	-1.20	0.78	0.44
	Upper HPD	0.40	2.53	2.47
CLR	Mean	-0.41	1.37	1.31
	s.e.	0.50	0.49	0.56
	Lower CL	-0.57	0.41	0.21
	Upper CL	1.39	2.33	2.41

Here “Mean” is the posterior mean, “s.e.” is the posterior standard deviation, “Lower HPD” and “Upper HPD” are the lower and upper end of the HPD region respectively, “Lower CL” and “Upper CL” are the lower and upper end of the confidence limit respectively.

Table 5-4: Results of the simulation study for the associated exposure

Associated Exposures: Full data				
Method		β_1	β_{21}	β_{22}
AM	Mean	2.1486	1.0235	0.7138
	MSE	0.1151	0.1398	0.0281
IM	Mean	2.2420	0.9370	0.7046
	MSE	0.2479	0.2543	0.1012
CLR	Mean	2.3129	1.1077	0.7383
	MSE	0.2171	0.2251	0.1682
Associated Exposures: Missing Data				
Method		β_1	β_{21}	β_{22}
AM	Mean	2.3689	1.0939	0.6309
	MSE	0.1628	0.1395	0.1194
IM	Mean	2.4614	1.0000	0.8824
	MSE	0.1828	0.4813	0.1295
CLR	Mean	2.4409	1.2744	0.8069
	MSE	0.4824	0.8018	0.5507

Results of the simulation study when the data was generated with association among exposures with $\lambda_{ij} = \gamma_{03i} + 0.44Z_{ij}$, $\gamma_{03i} \sim \text{Normal}(-1.40, 2.38)$. Here "Mean" is the mean of the 50 estimates corresponding to the 50 simulated datasets while MSE is the mean squared error. The true parameter values are $\beta_1 = 2.18$, $\beta_{21} = 1.11$ and $\beta_{22} = 0.67$. AM and IM stand for association and independence modeling respectively. Missing data was created by randomly deleting 20% of observations on $X^{(1)}$ and $X^{(2)}$.

Table 5-5: Results of the simulation study for the independent exposures

Independent Exposures: Full data				
Method		β_1	β_{21}	β_{22}
AM	Mean	2.2769	0.7913	0.7700
	MSE	0.2289	0.1200	0.0986
IM	Mean	2.4165	0.9710	0.7609
	MSE	0.1315	0.1647	0.0496
CLR	Mean	2.3919	1.2365	0.6647
	MSE	0.3277	0.2097	0.1960
Independent Exposures: Missing Data				
Method		β_1	β_{21}	β_{22}
AM	Mean	2.3165	0.8186	0.8883
	MSE	0.1426	0.3189	0.2380
IM	Mean	2.4955	0.9136	0.6286
	MSE	0.1812	0.2433	0.2162
CLR	Mean	2.4559	1.3303	0.8878
	MSE	0.4908	0.5052	0.6597

Results of the simulation study when the data was generated with independent exposures with log-odds ratio $\lambda_{ij} = 0.00$. Here “Mean” is the mean of the 50 estimates corresponding to the 50 simulated datasets while MSE is the mean squared error. The true parameter values are $\beta_1 = 2.18$, $\beta_{21} = 1.11$ and $\beta_{22} = 0.67$. AM and IM stand for association and independence modeling respectively. Missing data was created by randomly deleting 20% of observations on $X^{(1)}$ and $X^{(2)}$.

CHAPTER 6

SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

Case control studies have drawn considerable attention from epidemiologist as well as from statisticians. In this dissertation we considered the problem of analyzing matched case-control studies with partial missing observations in exposure variables assuming the data are missing at random. To handle the missingness I proposed a full likelihood based approach. The novelties of my approach lies in capturing the unmeasured stratum heterogeneity on the distribution of the exposure variables. I estimate the model parameters in a semiparametric Bayesian framework. This is the first unified Bayesian approach in matched case-control studies which can handle multiple disease states, missing exposure variables and the measurement errors in the exposure variables. I have also considered modeling association among multiple exposures in a matched case-control setup when some of the exposure variables are partially missing.

There are many possible extensions of my proposed methodology for analyzing case-control data. First, one can extend the proposed method of handling missing data to nested case-control, case-cohort (Prentice 1986), and population-based case-control study (Benichou and Gail, 1995; Benichou and Wacholder, 1994). Second, in my research I assumed presence of a completely observed covariate(s) in the datasets and modeled the distribution of the exposure variables in terms of the former. One can also investigate how to handle the problem of missing exposure in the absence of such completely observed covariate(s). Chatterjee et al. (2003) considered a similar problem for unmatched case-control studies. They used a semiparametric method following the approach of Lawless et al. (1999). But the analogous approach in the context of matched case-control studies is still unknown.

Another potential problem is to handle nonignorable missing values in a exposure variable in a case-control studies. Chapter 3 deals with missing exposure variable in matched case-control studies with multiple disease states. Disease can be classified on the basis of various pathological information on a subject. But all pathological information may not be observed for all the study subjects and consequently one faces the problem of missing information on the disease states which is not yet addressed in the literature.

Another possible area of research is to adapt my methods for analyzing survival data. Some work in this regard has been initiated in a frequentist framework by Prentice and Breslow (1978), but Bayesian inference for such problems seems to be largely unexplored. For example the methods could potentially be adapted to bivariate survival models in family based study design (see Oakes 1986, 1989 and Shih and Chatterjee, 2002). Finally, genetic case-control study is a new emerging area of research where one of the objectives is to study the association between a candidate gene and a disease. Accounting for population substructure through varying stratum effects is of particular relevance in such studies.

Bayesian approach could also be useful to analyze data coming from other designs used in an epidemiologic context (Khoury, Beaty and Cohen, 1993). For example, another frequently used design is a nested case-control design (Wacholder 1996) where a case-control study is nested within a cohort study. This type of study is useful to reduce the cost and labor involved in collecting the data on all individuals in the cohort as well as to reduce computational burden associated with time-dependent explanatory variable. Unlike the case-control design this design allows us to estimate the absolute risk of the disease by knowing the crude disease rate from the cohort study. Some other study designs along this line are the proband design (Gail, Pee and Carroll, 2001) and case-only design (Armstrong 2003).

It is important to note that case-control designs are choice-based sampling designs in which the population is stratified on the values of the response variable itself. Among others, this was noticed in Scott and Wild (1986) who compared in such cases a maximum likelihood based approach with some other ad hoc methods of estimation. Breslow and Cain (1988) considered in such cases a two-phase sampling design. In later years, there is a long series of publications on inference for such designs, but any Bayesian approach to these problems is still lacking, and will be a worthwhile future undertaking.

REFERENCES

- Altham, P. M. E. (1969). Exact bayesian analysis of a 2×2 contingency table and fisher's exact significance test, *Journal of Royal Statistical Society, B* **31**: 261–269.
- Anderson, J. A. (1972). Separate sample logistic discrimination, *Biometrika* **59**: 19–35.
- Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to non-parametric problems, *The Annals of Statistics* **2**: 1152–1174.
- Armitage, P. and Berry, G. (1994). *Statistical methods in medical research*, Blackwell Scientific Publications Ltd.
- Armstrong, B. G. (2003). Fixed factors that modify the effects of time-varying factors: Applying the case-only approach, *Epidemiology* **14**: 467–472.
- Armstrong, B. G., Whittemore, A. S. and Howe, G. R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer, *Statistics in Medicine* **8**: 1151–1163.
- Ashby, D., Hutton, J. L. and McGee, M. A. (1993). Simple bayesian analyses for case-controlled studies in cancer epidemiology, *Statistician* **42**: 385–389.
- Benichou, J. and Gail, M. (1995). Methods of inference for estimates of absolute risk derived from population-based case-control studies source, *Biometrics* **51**: 182–194.
- Benichou, J. and Wacholder, S. (1994). A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control study, *Statistics in Medicine* **13**: 651–661.
- Breslow, N. E. (1996). Statistics in epidemiology, the case-control study, *Journal of the American Statistical Association* **91**: 14–28.
- Breslow, N. E. and Cain (1988). Logistic regression for two-stage case-control data, *Biometrika* **75**: 11–20.
- Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research Volume I: The Analysis of Case-Control Studies*, Oxford University Press.

- Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies, *American Journal of Epidemiology* **108**: 299–307.
- Carroll, R. J. and Stefanski, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses, *Statistics in Medicine* **13**: 1265–1282.
- Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models, *Journal of the Royal Statistical Society, Series B, Methodological* **53**: 573–585.
- Carroll, R. J., Gail, M. H. and Lubin, J. H. (1993). Case-control studies with errors in covariates, *Journal of the American Statistical Association* **88**: 185–199.
- Carroll, R. J., Wang, S. and Wang, C. Y. (1995). Prospective analysis of logistic case-control studies, *Journal of the American Statistical Association* **90**: 157–169.
- Chatterjee, N., Chen, Y. H. and Breslow, N. E. (2003). A pseudo-score estimator for regression problems with two-phase sampling, *Journal of the American Statistical Association* **98**: 158–168.
- Cohen, N. D. (1997). Epidemiology of colic, *Veterinary Clinic North America Equine Practice* **13**: 91–201.
- Cornfield, J. (1951). A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix, *Journal of the National Cancer Institute* **11**: 1269–1275.
- Cornfield, J., Gordon, T. and Smith, W. W. (1961). Quantal response curves for experimentally uncontrolled variables, *Bulletin of the International Statistical Institute* **38**: 97–115.
- Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, Chapman and Hall Ltd.
- Dahm, P. F., Gail, M. H., Rosenberg, P. S. and Pee, D. (1995). Determining the value of additional surrogate exposure data for improving the estimate of an odds ratio, *Statistics in Medicine* **14**: 2581–2598.
- Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant, *Biometrics* **23**: 313–323.
- Diggle, P. J., Morris, S. E. and Wakefield, J. C. (2000). Point-source modelling using matched case-control data, *Biostatistics* **1**(1): 89–105.
- Ekholm, A., McDonald, J. W. and Smith, P. W. F. (2000). Association models for a multivariate binary response, *Biometrics* **56**(3): 712–718.

- Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior, *Journal of the American Statistical Association* **89**: 268–277.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**: 577–588.
- Forbes, A. B. and Santner, T. J. (1995). Estimators of odds ratio regression parameters in matched case-control studies with covariate measurement error, *Journal of the American Statistical Association* **90**: 1075–1084.
- Fuller, W. A. (1987). *Measurement Error Models*, John Wiley and Sons.
- Gail, M. H., Pee, D. and Carroll, R. (2001). Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies, *Journal of Statistical Planning and Inference* **96**(1): 167–177.
- Gelman, A. and Rubin, D. B. (1992). Reply to comments on “Inference from iterative simulation using multiple sequences”, *Statistical Science* **7**: 503–511.
- Ghosh, M. and Chen, M.-H. (2002). Bayesian inference for matched case-control studies, *Sankhya, Series B, Indian Journal of Statistics* **64**(2): 107–127.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, John Wiley and Sons.
- Khoury, M. J., Beaty, T. H. and Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*, Oxford University Press.
- Kim, I., Cohen, N. D. and Carroll, R. J. (2002). A method for graphical representation of effect heterogeneity by a matched covariate in matched case-control studies exemplified using data from a study of colic in horses, *American Journal of Epidemiology* **156**: 463–470.
- Lang, J. B., McDonald, J. W. and Smith, P. W. F. (1999). Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach, *Journal of the American Statistical Association* **94**: 1161–1171.
- Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression, *Journal of the Royal Statistical Society, Series B, Methodological* **61**: 413–438.
- Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association* **88**: 1341–1349.
- Lipsitz, S. R., Parzen, M. and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates, *Biometrics* **54**: 295–303.

- Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, John Wiley and Sons.
- MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models, *Journal of Computational and Graphical Statistics* **7**: 223–238.
- Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**: 719–748.
- Marshall, R. J. (1988). Bayesian analysis of case-control studies, *Statistics in Medicine* **7**: 1223–1230.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**: 153–157.
- McShane, L. M., Midthune, D. N., Dorgan, J. F., Freedman, L. S. and Carroll, R. J. (2001). Covariate measurement error adjustment for matched case-control studies, *Biometrics* **57**(1): 62–73.
- Miller, A. B., Howe, G. R., Jain, M., Craib, K. J. P. and Harrison, L. (1983). Food items and food groups as risk factors in a case-control study of diet and colo-rectal cancer, *International Journal of Cancer* **32**: 155–161.
- Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables, *Biometrika* **84**: 523–537.
- Müller, P., Parmigiani, G., Schildkraut, J. and Tardella, L. (1999). A Bayesian hierarchical approach for combining case-control and prospective studies, *Biometrics* **55**: 858–866.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics* **9**(2): 249–265.
- Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions, *Scandinavian Journal of Statistics* **14**: 67–77.
- Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data, *Biometrika* **73**: 353–361.
- Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association* **84**: 487–493.
- Paik, M. C. and Sacco, R. L. (2000). Matched case-control data analyses with missing covariates, *Applied Statistics* **49**(1): 145–156.
- Pastides, H., Kelsey, J. L., Holford, T. R. and LiVolsi, V. A. (1985). An epidemiologic study of fibrocystic breast disease with reference to ductal epithelial atypia, *American Journal of Epidemiology* **121**: 440–447.

- Phillips, A. and Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate, *Biometrics* **43**: 425–431.
- Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model (Corr: V71 p219), *Biometrika* **69**: 331–342.
- Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**: 1–11.
- Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models, *Biometrika* **65**: 153–158.
- Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* **66**: 403–412.
- Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies, *Journal of the Royal Statistical Society, Series B, Methodological* **65**: 711–723.
- Rathouz, P. J., Satten, G. A. and Carroll, R. J. (2002). Semiparametric inference in matched case-control studies with missing covariate data, *Biometrika* **89**(4): 905–916.
- Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag Inc.
- Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models, *Biometrics* **42**: 311–323.
- Satten, G. A. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates, *Biometrics* **56**(2): 384–388.
- Satten, G. A. and Kupper, L. L. (1993a). Conditional regression analysis of the exposure-disease odds ratio using known probability-of-exposure values, *Biometrics* **49**: 429–440.
- Satten, G. A. and Kupper, L. L. (1993b). Inferences about exposure-disease associations using probability-of-exposure information, *Journal of the American Statistical Association* **88**: 200–208.
- Satten, G. A., Flanders, W. D. and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model, *American Journal of Human Genetics* **68**: 466–477.
- Schottenfeld, D. and Fraumeni, Joseph F., J. E. (1996). *Cancer Epidemiology and Prevention (Second edition)*, Oxford University Press.

- Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling, *Journal of the Royal Statistical Society, Series B, Methodological* **48**: 170–182.
- Seaman, S. R. and Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates, *Biometrika* **88**(4): 1073–1088.
- Seaman, S. R. and Richardson, S. (2004). Equivalence of prospective and retrospective models in the bayesian analysis of case-control studies, *Biometrika* **91**: 15–25.
- Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case-control family studies, *Biometrics* **58**(3): 502–509.
- Sinha, S., Mukherjee, B. and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states, *Biometrics* **60**: 41–49.
- Sinha, S., Mukherjee, B., Mallick, B. K., Ghosh, M. and Carroll, R. (2003). Semiparametric bayesian analysis of matched case-control studies with missing exposure, *Revised and Re-submitted*.
- Wacholder, S. (1996). The case-control study as data missing by design: Estimating risk difference, *Epidemiology* **7**: 144–150.
- Walsh, R. D. (1994). Effects of maternal smoking on adverse pregnancy outcomes: Examination of the criteria of causation, *Human Biology* **66**: 1059–1092.
- Wang, C. Y., Wang, S. and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis, *Journal of Econometrics* **77**: 65–86.
- West, M., Mller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation, *Aspects of Uncertainty. A Tribute to D. V. Lindley*, pp. 363–386.
- Whittemore, A. S. (1991). Errors-in-variables regression problems in epidemiology, *Statistical Analysis of Measurement Error Models and Applications*, pp. 17–31.
- Zelen, M. and Parker, R. A. (1986). Case-control studies and Bayesian inference, *Statistics in Medicine* **5**: 261–269.
- Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**: 642–648.

BIOGRAPHICAL SKETCH

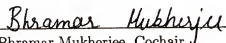
Samiran Sinha, (nickname, Rana) was born on October 25, 1976, in Naihati (a small town, a few miles away from Calcutta, India). He received his undergraduate degree in 1997 from Kalyani University (with honors in statistics, and minoring in mathematics and physics). He received his master's degree in statistics (with specialization in advanced design of experiments, operation research, and optimization technique) from Calcutta University in 1999.

He joined the graduate program in statistics at the University of Florida in August 2001. He passed his master's comprehensive examination and Ph.D qualifying examination in May, 2002 and August, 2002 respectively. He started to work on his dissertation in the Summer of 2002. From August 2004, he will be on the faculty of the Department of Statistics at Texas A&M University as an assistant professor.

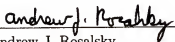
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


Malay Ghosh, Chair
Distinguished Professor of Statistics

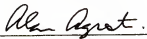
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


Bhramar Mukherjee, Cochair
Assistant Professor of Statistics

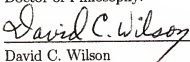
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


Andrew J. Rosalsky
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

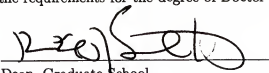

Alan G. Agresti
Distinguished Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


David C. Wilson
Professor of Mathematics

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 2004



Dean, Graduate School